DIGITAL ARCHIVING AT NASA GODDARD SPACE FLIGHT CENTER

Part 1: Benchmarks and Current Activities

NASA GSFC Library

Final Draft
September 30, 2002

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Goddard Space Flight Center (GSFC) relies increasingly on electronic means to record and disseminate information about its missions, activities and operations. The mission of the Library is to preserve and provide access to the knowledge assets needed to carry out the Center's mission, and preservation of the GSFC's digital assets are a natural outgrowth of the GSFC Library's current activities. Therefore, the GSFC Library conducted a project to evaluate the environment for digital preservation at GSFC and to develop a framework for preserving GSFC digital assets.

Two types of resources were emphasized during this pilot project. The first is the video capture of colloquia, mini-courses, and other activities in which internal and external experts share the results and lessons learned from projects. This has been identified as a key component of the GSFC Knowledge Management initiative. The second pilot captured Web sites, with an emphasis on project Web sites and those from directorates that conduct scientific and technical research and engineering activities.

This report is provided in two parts. Part 1 defines digital archiving and preservation and the key challenges involved; describes external digital preservation projects that apply to the GSFC environment; and reports on the two pilot projects described above. Part 1 is intended to assess the current state of digital archiving and preservation, the current state of the practice, and relevant standards and guidelines. Part 2 provides a framework for moving GSFC and the GSFC Library from the current state closer to an infrastructure for archiving and preservation. The framework does not provide all the answers but provide a methodology and infrastructure within which GSFC's Library, the owners of critical content, potential users, and other key stakeholder groups can work together to achieve a working system.

The video capture project has successfully been implemented. The GSFC Library is now capturing and providing access to most of the colloquia, mini-courses, and other similar activities that occur on Center. Proposals have been made to expand this service and to provide portable equipment for use by others with the videos or webcasts being encoded and stored at the GSFC Library. Minimal metadata is being supplied at this time. Additional efforts are underway to convert speech to text in order to provide indexing for the content. The team has also successfully provided simultaneous access to the video of the speaker and the capture of the PPT slides. The slides have also been converted into text in order to provide additional indexing.

The Web capture pilot involved several key subtasks. It began with an analysis of a sample of the GSFC project and non-project Web sites. They were characterized by the number and types of digital objects contained on the pages, the complexity of the Web design, their audience, and number and types of links. Selected project and non-project sites with scientific and technical content were used for several test capture runs using a freeware spidering tool called HTTrack. This identified several issues, including the number of levels to be used in the spidering, how to handle linked sites that are not in the GSFC domain, intellectual property issues, and complex and large linked objects such as video files and data sets, and broken links and inaccessible pages. Problems such as viruses and speed of processing identified the need for an isolated computer system to capture, virus check and scrub the captured data. The size of the resulting files may be prohibitive and while significant storage will be needed there are outstanding issues

related to compression and near-line versus online availability. Additional analysis and policy decisions are needed to balance the benefits of complete capture of the sites with the resources available.

Based on these pilot projects, the analysis of the current state of best practices in digital archiving and preservation, and the understanding of the GSFC environment, the team outlined in Part 2 an infrastructure. This infrastructure supports ongoing development of a digital preservation strategy for GSFC, particularly as it relates to information from projects.

First, the team recommends the creation of a GSFC Digital Preservation Steering Committee composed of project librarians, managers from the GSFC Library, potential users, people involved in KM initiatives, and other stakeholder groups. This group will review the outstanding issues that are identified in Part 2 of this report and work to tailor the needs for digital preservation to the GSFC environment and resources. A high level champion must be found for this group. The group would be supported by the GSFC Library and its contractor staff.

At the conceptual level of the infrastructure, the team recommends that the Open Archival Information System Reference Model (ISO 1472) should serve as the framework for the digital preservation system at GSFC. Even before its recent adoption as an ISO standard, OAIS was used by all the major digital preservation activities being conducted by national libraries, archives, and special collections. It provides definitions, a data model and a functional model into which the key metadata packets can be plugged. Expertise regarding the OAIS is readily available at GSFC, since Donald Sawyer of Code 630 was the NASA representative and spearheaded the development of the model.

As part of the overall infrastructure analysis, the project analyzed current metadata schemes applicable to the digital objects and the subject matter of importance to the GSFC community. Based on the analysis of the types of digital objects included in GSFC project Web sites, a number of different metadata schemes would be applicable. Also, the team investigated metadata of importance for preservation and specialized sets required for geospatial referencing and digital still image content. The team recommends the minimal set of Dublin Core metadata, with the opportunity in the future to extend the set to include elements of importance to projects.

In addition the Metadata Encoding and Transmission Standard (METS) was investigated through a joint project with the College of Information Studies at the University of Maryland, College Park, to determine if this metadata framework is applicable. It was determined that METS could provide a framework in which the project files could be managed. Key implementation questions were identified for follow-on activities.

As a result of the pilot projects and the investigation of the state of the art and practice of digital preservation, the team has developed a high level conceptual design and has created some preliminary programs to prototype a semi-automated production system for Web capture. The text captured from the Web sites has also been indexed using Autonomy Server, a search engine already licensed by the GSFC Library and other organizations at GSFC. The results of the Web capture are available from a default Autonomy interface.

1.0 PURPOSE AND BACKGROUND OF THE PROJECT

Goddard Space Flight Center (GSFC) relies increasingly on electronic means to record and disseminate information about its missions, activities and operations. The mission of the Library is to preserve and provide access to the knowledge assets needed to carry out the Center's mission, and preservation of the GSFC's digital assets are a natural outgrowth of the GSFC Library's current activities. Coordination and planning for digital preservation and long-term access are key to the provision of content, a critical infrastructure component for knowledge management. Therefore, the GSFC Library conducted a project to evaluate the environment for digital preservation at GSFC and to develop a framework for preserving GSFC digital assets.

This report is in two parts. Part 1 defines digital archiving and preservation and the key challenges involved; describes the national and international projects and best practices applicable to the issues that GSFC faces; and reports on the current situation at GSFC, including the results of two pilot projects conducted by the GSFC Library. Part 1 is intended to assess the current state of digital archiving and preservation, the relevant benchmarks and state of the practice, and relevant standards and guidelines.

Part 2 provides a framework for moving GSFC and the GSFC Library from the current state to a state closer to the benchmarks. It provides several guiding principles, and an implementation plan with proposed activities, priorities, resources and timelines. The framework does not provide all the answers but seeks to provide a methodology and infrastructure within which GSFC's Library, the owners of critical content, potential users, and other key stakeholder groups can work together to achieve a working system. Because the plan would require commitments throughout the Center, the framework addresses the resources, training and social/cultural issues involved.

2.0 DIGITAL ENVIRONMENT: CHALLENGES AND OPPORTUNITIES

Digital information can be born digital or digitized from analog (a printed text, photograph, map, etc.). Materials that are born digital are those that were created in a digital environment or materials whose major preservation format is the digital form. Materials that are born digital do not have the analog version as a backup. While this report focuses on materials that are born digital, the same principles generally apply if the analog version is superseded by the digital version once the analog version has been digitized.

In many ways the digital environment is more fragile than the paper environment. This fragility comes from the close coupling of the technology to the content. In cases, such as multimedia, simulations, or game technologies, it is almost impossible to separate the content from the machine used to display/run it. Digital objects can be changed intentionally or unintentionally with little recognition that the change has occurred.

The ease with which digital materials can be created (almost everyone can be an author or a publisher on the Web) has caused a dramatic increase in what could be archived. It has also taken the publishing and, ultimately, the archiving of these materials outside the primary life

cycle (author to publisher to library to archive), which has supported the print environment, particularly in the sciences, for over a century.

Digital archiving and digital preservation are often used interchangeably. However, there is actually a significant difference between the two concepts. In this report, "digital archiving" is defined as a specific event or point in time when a digital object, whether born digital or converted to digital form from analog, is stored. "Digital preservation" is an ongoing activity necessitated by the ever-changing technologies involved in the digital environment. Preservation requires ongoing planning, decision-making, and stewardship.

An adjunct to preservation is long-term access. It is important to note that few, if any, projects have put any parameters on what is meant by long-term. Long-term is often defined as the length of time the material would be of value to a particular community [CCSDS, 2002]. Long-term access is extremely important, since preserving something that cannot be reused is both expensive and foolhardy. Unfortunately, long-term access is the most difficult aspect of digital preservation, since it relies on access technologies.

The differences between the analog and digital cultures, the new stakeholder groups involved in digital publishing including IT professionals, and the rapid technological changes have raised serious challenges for digital archiving and preservation. Unlike the preservation of non-digital materials, which has standards and well-established institutions and support services, the digital environment is just beginning to develop an infrastructure. However, just as in the print environment, it is important for an institution such as GSFC to look for guidance externally, but to implement locally. While not all the answers are available at this point in the development of the body of knowledge regarding digital archiving and preservation, there is sufficient consensus to identify best practices to inform local implementations, and the penalty for waiting is even greater. In addition, the digital environment provides unique opportunities for improving the scientific and engineering communication chain that it is imperative to move forward even in this uncertain environment.

3.0 THE GODDARD ENVIRONMENT

Key aspects of the environment include GSFC's knowledge management initiatives, which serve as the context for this effort, the content (object types and formats) of importance to GSFC's initiatives, and existing preservation projects that could be identified.

3.1 Goddard's Knowledge Management Initiatives

The draft Knowledge Management Strategic Plan provides the most comprehensive vision for GSFC's knowledge management environment. It focuses not only on the technologies that will support knowledge collection, management and dissemination, such as the MyGoddard portal, but on the institutional culture and social infrastructure (incentives and rewards) that are of equal or greater importance than the technologies. The preservation of a variety of digital types is critical to the success of this vision.

In addition to the overall knowledge management planning that is underway at GSFC, there are several related projects with which the Library has been involved. These include the EOS Lessons Learned Pilot, the CIO Pilot to Automatically Categorize Project Documents, and the Multimedia Asset Management System investigation being conducted by the Knowledge Management Officer.

While the knowledge management strategy does not specifically address long-term preservation, this is essential to ensure that information of value is not lost when project funding ends. In addition to simply providing a place to store this information, there must be ongoing stewardship of the information and provision of access mechanisms over time.

The Library is well positioned to serve as a key player in the development of the knowledge management culture at GSFC [GSFC Library Visiting Committee Report, 2002]. Digital archiving and preservation are natural extensions of the Library's current mission to identify, select, organize, and provide access tailored to the needs of the GSFC mission. The Library has expertise in archiving and preserving internal and external print materials of importance to GSFC's mission and in providing access to an increasing amount of electronic material.

3.2 Goddard's Object Types

Before a system can be defined that will support the GSFC knowledge management objectives, it is important to analyze the types of content that are of importance to the GSFC environment.

The following major content types were identified based on external projects and an analysis of the GSFC environment. This is not a comprehensive list, and, in practice, there are overlaps between these categories.

Fact Sheets – brief descriptions of missions, technical developments or other outcomes, many of which are designed to promote technology transfer to industry

Project Documentation – including, but not limited to the information required by the NASA Guidelines and outlined in the NODIS system.

Outreach materials – materials prepared for schools, the public, journalists, etc., which may include press releases, curriculum packets, and public web sites

Published journal articles and books – report of scientific and technical discoveries and designs that are formally disseminated through commercial or not-for-profit publishers

Presentations – the speech, text and handouts used to give oral reports of scientific and technical work at formal or informal meetings, colloquia, training sessions, conferences, etc.

Technical reports and conference proceedings – reports of work that are more informally published, usually by NASA or the sponsoring organization, in which the publication process does not involve full peer review and dissemination is through more informal and less market-driven channels

The content types outlined above may appear in many object types including the following, which were identified from an analysis of the GSFC project and non-project materials. The following categorization is loosely based on a scheme developed for the British Library [Hendley, 1998].

Audio - voice recordings that are not included as part of video

Data – numeric or alphanumeric data sets often created by instruments, laboratory equipment, or computer. The software to manipulate or visualize the data is under Software.

Software/Simulations and Other Application Tools - a variety of application programs, including software tools for manipulating and analyzing data.

Still Images - two-dimensional fixed images such as photographs or digitized TIFF images of documents, maps, or other textual materials. It does not include three-dimensional images.

Text –word-based materials, including books, journal articles, manuscripts, reports, technical reports, project documentation, etc.

Video – full motion pictures, mostly with sound included

Web sites per se are not included in this list of object types since the Web is a publication and dissemination medium, which can convey any or all of the above content and object types. While it is the medium of choice at this time, it is impossible to say what will be the "standard" in 20 or 30 years. However, the medium of the Web presents certain opportunities and challenges with regard to digital preservation. First, as stated before the Web makes publishing easy and, therefore, the rigors of previous publishing environments, even internal ones, can be more easily bypassed, requiring new kinds of structures and policies to ensure adequate preservation. On the other hand, Web access to information allows preservation groups to capture the pages to which they have access with little effort on the part of the creator.

An analysis of a sample of project sites on the GSFC Web domain identified the following object types by year.

Year	Audio	Data set Links	Software	Still Images	Image links	Text	Video	Video Links
2001	0	Yes	1	9	Yes	1	0	0
2000	0	0	0	3	Yes	14	0	0
1999	0	0	0	16	Yes	27	0	0
1998	0	Yes	0	3	Yes	15	0	0
1997	0	Yes	1	2	Yes	19	0	
1996	0	Yes	0	5	Yes	10	0	0
1995	0	Yes	1	5	Yes	54	0	
1994	0	Yes	0	2	Yes	17	0	0
1993*	0	Yes	0	1	Yes	7	0	0
1992*	0	Yes	0	1	Yes	7	0	0
1991*	0	Yes	0	1	Yes	10	0	0
1990	0	Yes	1	1	Yes	82	0	0
1989(1^{st})	0	Yes	1	1	Yes	46	0	0

1988(1st)	0	0	0	1	Yes	7	0	0
1987(1st)	--	--	--	--	--	--	--	--
1986*	0	Yes	0	1	Yes	5	0	0
1985*(1st)	0	Yes	0	1	Yes	5	0	0
1984	0	Yes	0	1	Yes	20	0	0
1983*	0	Yes	0	1	Yes	8	0	0

(*) For the fifth project of this particular year, only a project PDF was given. (NSSDC Master Catalog entry). The PDF column was labeled as "1" because the page is a .PDF.

(1st) All entries are from the second survey of the project sites, where every fifth project site was sampled per year, unless noted as from the first survey by (1st), where every first site was sampled per year. Reasons for including site from the first survey include: only one working site for given year, all sites for that year had already been sampled, etc.

(--) The Website from 1987 originally surveyed is now a 404.

It is clear that the Web sites have become increasingly complex in terms of the types of objects included. (Examples of GSFC project and nonproject homepages are included in Appendix H.) Where as in the early projects the emphasis was on text, Web sites now contain additional object types in more complex Web page designs. There is an increase in the number and complexity of graphics. There are full motion video clips, sound bytes and visualizations that were not previously included. They may have been available on CD-ROMs or local servers, but the state of Web technology did not allow their incorporation or easy access via standard Web browsers.

With the inclusion of a wider variety of digital objects, the types of file formats have also increased.

Year	.PDF	.JPEG	.DOC	.GIF	.MPEG	.XLS	.RTF	HTML/Other	Email
2001	0	13	0	0	0	0	0	1	1
2000	0	1	0	0	0	0	0	14	1
1999	0	8	0	0	0	0	0	27	2
1998	0	2	0	3	0	0	0	15	1
1997	0	1	0	41	0	0	0	19	1
1996	0	4	0	1	0	0	0	10	2
1995	0	3	0	3	0	0	0	54	1
1994	0	0	0	19	0	0	0	17	1
1993*	1	0	0	1	0	0	0	7	1
1992*	1	0	0	1	0	0	0	7	1
1991*	1	0	0	1	0	0	0	10	1
1990	0	0	0	4	0	0	0	82	0
1989(1st)	0	1	0	37	0	0	0	46	1
1988(1st)	0	1	0	12	0	0	0	7	1
1987(1st)	--	--	--	--	--	--	--	--	--
1986*	1	0	0	1	0	0	0	5	1
1985*(1st)	1	0	0	0	0	0	0	5	1
1984	0	0	0	1	0	0	0	20	2
1983*	1	0	0	1	0	0	0	8	1

(*) For the fifth project of this particular year, only a project PDF was given. (NSSDC Master Catalog entry). The PDF column was labeled as "1" because the page is a .PDF.

(1st) All entries are from the second survey of the project sites, where every fifth project site was sampled per year, unless noted as from the first survey by (1st), where every first site was sampled per year. Reasons for including site from the first survey include: only one working site for given year, all sites for that year had already been sampled, etc.

(--) The Website from 1987 originally surveyed is now a 404.

3.3 Existing Projects at Goddard

In addition to the work done in this area by the GSFC Library, several other archiving and preservation projects have been identified at GSFC. It is difficult to get a comprehensive list of these types of projects, so the following are only examples. The key point is that GSFC is not starting from scratch with regard to these issues because the lessons learned from other activities are relevant to digital preservation issues.

3.3.1 NASA and the Open Archival Information System Reference Model

As a member of the Consultative Committee on Space Data Systems, NASA, through representatives from GSFC, has had a lead role in the development of the Open Archival Information Systems Reference Model. This model provides a high-level data and functional model for digital (and physical) archives. The OAIS RM was developed in response to the request by the ISO Technical committee to develop standards for digital archiving. In the end, the RM proved to be generalizable across many object types and sectors, including libraries and records management systems. In June the OAIS RM became ISO Standard 1472. The NSSDC archive is considering the use of the OAIS Model. The OAIS Model is described in greater detail in section 4.7.

3.3.2 NASA DAACs

The NASA DAACs are some of the oldest archives. This "network" of subject oriented data collection centers archives and preserves data sets from NASA and other agencies. The indivdual DAACs form a loosely federated network, and collectively are part of the World Data Centers. In addition to data management, the DAACs are involved in providing access and information products based on the data sets for which they have stewardship. Many of their customers download the data, though other mechanisms such as ftp and CD-ROM are also used.

Much of the data archived by the DAACs is not extremely complex, but there is a large volume of numeric and alphanumeric data that must be managed. One of the crucial issues related to long-term preservation of data is the slow speed of input-output devices compared to the speed of machines and the cost of storage. Storing the massive amount of information is not a problem, but it is a problem to ensure its preservation by refreshing the media. One data center indicated that it would not be long before they would not be able to complete one round of migration before they should begin another (assuming an 8-10 year replacement cycle) [Hodge, 1999].

This issue was recently reflected in the outcome of work by the Library of Congress and the National Science Foundation that brought the research community and government agencies together. The effort seeks to define a research agenda for digital archiving that will support the needs of federal government agencies as well as the LC's digital preservation infrastructure

activities for the nation [Committee on an Information Technology Strategy for the LC, 2001]. The results will be provided back to the government for its use and also to the private sector for further development. The major research areas identified to-date are the migration of extremely large data sets and long-term access to complex multimedia objects.

3.3.3 NASA Digital Television

[need more information about this]

4.0 STATE OF THE ART AND PRACTICE OUTSIDE OF GODDARD

Since the early 1990s, there have been projects involving digital archiving. While many of the early projects involved the preservation of cultural heritage information, several have included scientific information, particularly electronic journals. Guidelines, best practices and lessons have been learned and shared. While the need to raise awareness about the importance of digital preservation has not disappeared, more time and words are being spent on the testing and implementation of pragmatic digital preservation projects, and the focus of research and development has shifted to "filling in the gaps."

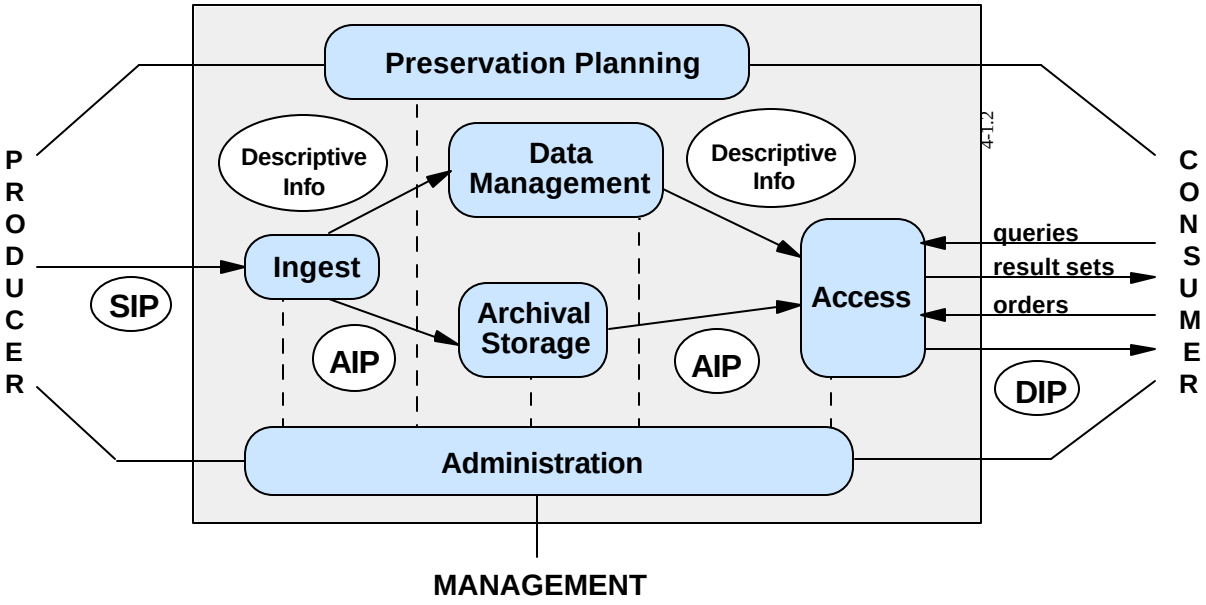
Until recently, many of the guidelines and best practices were narrowly defined in the context of each organization's local needs. During the last two years, there has been significant movement toward the identification of more broadly applicable best practices and standards. The Project has finalized its key guidance documents [Cedars, 2002, April]. At the U.S. National Agricultural Library, guidelines and a template for metadata capture for USDA digital publications are in the final approval stage [National Agricultural Library, 2002]. *Preservation Management of Digital Materials: A Handbook*, a comprehensive look at the outcomes of all the major projects, was published by the British Library and the UK Joint Information Systems Committee [Jones & Beagrie, 2001].

Best practices and benchmark projects can be identified for the following areas: selection criteria, metadata for description and preservation, content mark-up, technical preservation strategies, transformation versus native formats, and organizational models for archiving.

4.1 A Reference Framework

The existence of an underlying framework or reference model for digital archiving has been a major factor in the advancement of digital preservation efforts. The Open Archival Information System Reference Model (OAIS RM) the origins of which were described in section 3.3.1 above, has become a cornerstone for digital preservation practices. The OAIS RM provides high level data and functional models and terminology that help stakeholder groups discuss digital preservation with a common frame of reference [CCSDS, 2002]. The OAIS RM has proven flexible enough to respond to communities as diverse as scientific data centers, national archives, cultural heritage institutions, and national libraries.

The OAIS RM describes key participants in preservation: the producer/creator, the archive, management and the customer. It defines the major information packages and the functions to be performed by a compliant archive.



(Used with permission from the Consultative Committee on Space Data Systems.)

SIP – Submission Information Packet (what is submitted or acquired from the producer)
AIP – Archival Information Packet (the object that is archived)
DIP – Dissemination Information Packet (the object that is distributed based on access requests)
Descriptive Info – metadata

Figure 1. OAIS Reference Model

In addition to the participants, the OAIS identifies high-level functions. Acquisition involves making arrangements with producers for receiving archive material. This may involve licensing or negotiations about the formats that should be used. Ingest is the act of bringing the content into the archive when the archive takes control of the material and creating standardized metadata as needed to support description and management. Data management and archival storage include the storage and routine media refreshment for the data and the metadata. Access provides search engines and finding aids for use by consumers. Administration handles the day to day provision of resources to support the archive's activities. Preservation planning sets the strategies for ongoing curation of and access to the archived objects and the metadata.

The OAIS RM became an ISO Standard 1472 in June 2002. All major preservation projects use the OAIS Reference Model, including the Electronic Records Archives Project at the U.S.

National Archives and Records Administration], the Library of Congress, the Cedars Project, the Networked European Deposit Library Project [NEDLIB, 2001], InterPARES [InterPARES, 2002] and OCLC's Digital Archive [OCLC Digital Archive, 2002]. A December 2001 symposium sponsored by CENDI, a group of senior scientific and technical information managers in the U.S. federal government, and the U.S. Federal Library and Information Center Committee, focused on the use of the OAIS RM as a bridge among the library, records management and Chief Information Officer communities [CENDI, 2001].

Along with the development of the framework led by NASA, other groups are extending the guidelines and infrastructure provided by the OAIS RM. The CNES in France has led the effort to further describe the interface between producers and the archive. The Research Libraries Group has taken the lead in the identification of a checklist to support certification of OAIS compliant archives. In August 2001, RLG produced a draft set of attributes for a trusted archive [Research Libraries Group, 2001]. A certification program would indicate compliance with digital archiving standards, such as the OAIS RM. Organizations that contract with certified archives would be assured of a particular level of information management and integration and portability with other OAIS-compliant archives. This checklist will also serve as a benchmark for the development of reliable internal archives such as that envisioned for GSFC. The draft report suggests an official certifying body to identify the attributes that would be measured, how assessments would be conducted and the procedures for revocation of an archive's certification. Of course many questions remain to be answered, including the identity of the certifying body.

While the OAIS RM is a high level framework, it is hard to put such a conceptual design into terms that are applicable to a specific archive. It is often difficult to see how the OAIS relates to the standards and best practices described above. However, the incorporation of standards and lessons learned from the various projects can be seen in an example from the U.S. National Agricultural Library (NAL). In this conceptual model of an archival system proposed for the preservation of the U.S. Department of Agriculture's electronic publications, the NAL has incorporated other standards within the OAIS RM framework. The other standards and methodologies include LOCKSS (Lots of Copies Keep Stuff Safe) for redundancy and validation, XML for interoperability of metadata, the extended CORC metadata for preservation, and a link to its traditional library cataloging system through a conversion to the MARC standard.

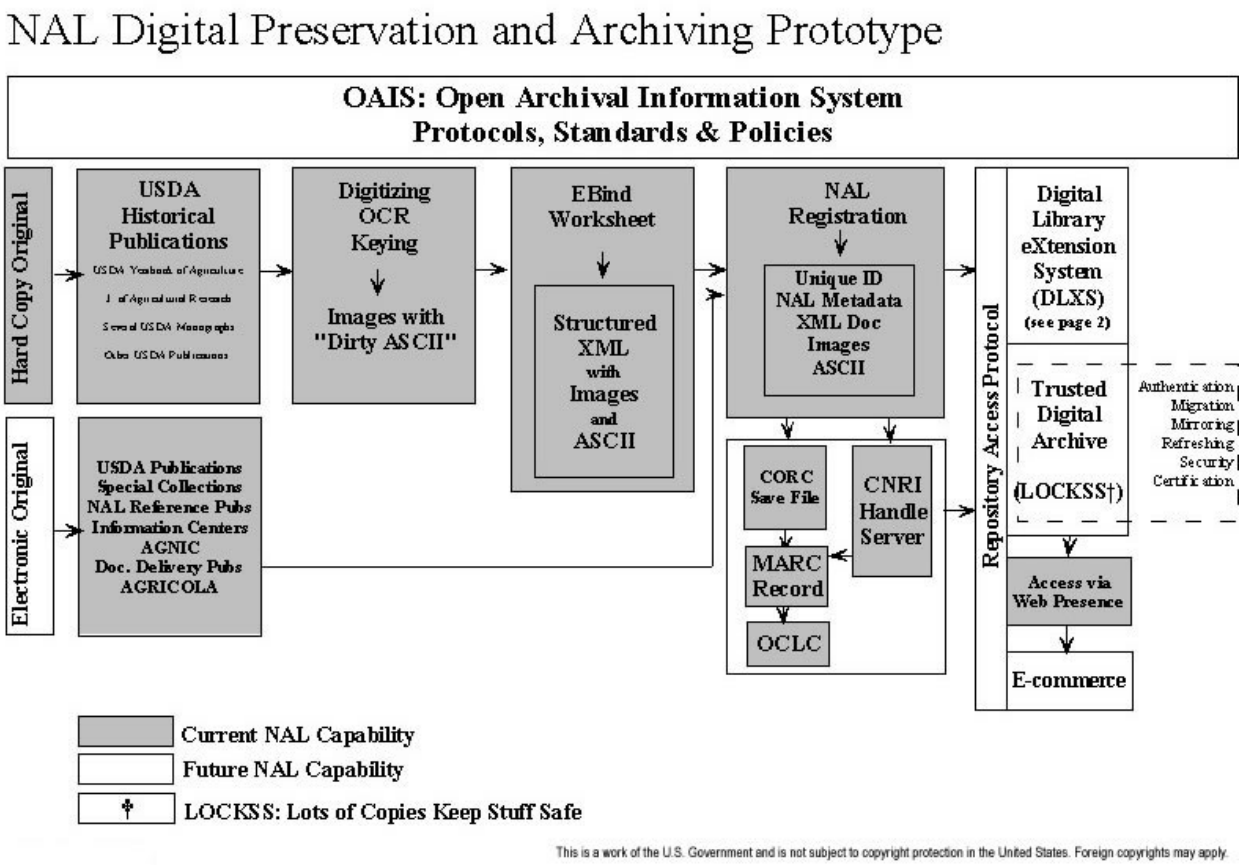


Figure 2: National Agricultural Library's Proposed Preservation Prototype

4.2 Collection Management

The Cedars Project was one of the earliest to deal with major issues regarding digital preservation. The conclusions based on over 3 years of studies and pilot projects is that the effective archives will be those that incorporate digital materials into the regular collection management and workflow, beginning with selection.

4.2.1 Selection Criteria

The key projects related to selection criteria were performed by Cedars [Cedars, 2002], the National Library of Australia [PANDORA, 2002], the National Library of Finland [Lounamaa & Salonharju, 1999] and the Royal Library of Sweden [Royal Library. National Library of Sweden, n.d.]. The Cedars project in particular emphasized the need for clear selection guidelines. The national libraries are dealing primarily with the capture of Web sites in cultural heritage. However, they learned many lessons regarding selection that can be transferred to other subject areas.

Issues that must be considered with regard to selection criteria include what sites are worth capturing and how long they should be retained, what portions of the site should be captured, what is the extent of a site, and how should links be managed. For example, the National Library

of Finland copies only Web sites within the same domain, regardless of the server. Others copy the site only if the link is to the same server as the homepage.

All these questions should be answered in the selection component of the Collection Management Guidelines to ensure agreement on what should be captured, to inform users about the archive's scope and contents, and to avoid legal problems.

4.2.2 Content Mark-up

Archiving requires storage of the content of digital objects in such a way that it can be rendered as needed for future use by a particular community or communities. This has led to research into standardized markup for particular document types, such as electronic journals and technical reports.

Harvard University has an archiving project underway with several major publishers including Elsevier, the American Institute of Physics and Nature, funded by the Andrew J. Mellon Foundation. A recent study analyzed the feasibility of a single SGML DTD (Standard Generalized Mark-up Language Document Type Definition) or XML schema for the deposit and archiving of electronic journals from different publishers [Inera, 2001].

With regard to technical reports, the technical report standard from the National Information Standards Organization (Z39.16) is under review. The Defense Technical Information Center of DoD has taken the lead in this review. As part of that activity, Old Dominion University has developed an XML schema for technical reports, which will provide easier access to the metadata as well as to the content.

4.3 Metadata

Because of the variety of object types identified in the earlier analysis in Section 3.2, there are a number of external metadata projects that are relevant to the GSFC environment.

4.3.1 A Metadata Framework

Because there are numerous metadata schemes that apply to GSFC a framework for incorporating the various schemes is of interest. The Metadata Encoding and Transmission Standard (METS) developed by the Library of Congress, provides such a framework. It divides metadata elements into several components including descriptive, structural and administrative. Within this structure and by using an XML schema, other standards can be referenced as needed. In addition, METS can be used, through its structural component, to hold a digital library collection together. For an analysis of the applicability of METS to the GSFC environment, see Part 2, Appendix B.

4.3.2 Descriptive Metadata

The most common standard for descriptive metadata is the Dublin Core (see Appendix A). This set of 15 elements provides a minimal set of information for resource discovery. It is used by

over 70 documented projects and is the basis for numerous other standards activities, such as the Open Archive Initiative, which is described in Part 1, Section 4.3.6.

The basic 15 elements can be enhanced by using the qualified Dublin Core instead of the unqualified. The qualified Dublin Core provides a mechanism for being more specific about certain fields. For example, by qualifying the meaning of the Date field, it is possible to set the date element to mean the creation date versus the publication date for each record.

The qualified Dublin Core provides the most flexibility giving a mechanism for incorporating other standards and other metadata elements. For example, it is possible to incorporate certain FGDC elements necessary by Presidential order for documenting objects that can be geospatially referenced by latitude and longitude coordinates.

The Dublin Core is the set that is most often mapped to other metadata schemes. For example, Dublin Core has been mapped to FGDC and to MARC, which is the metadata standard used in most library cataloging systems. This Dublin Core records for Web sites, for example, to be used along with more complete bibliographic records in MARC for other library resources.

4.3.3 Preservation Metadata

The Preservation Metadata Working Group of RLG/OCLC evaluated the state of the art in preservation metadata. The preservation metadata schemes developed by NEDLIB, the National Library of Australia, Cedars and the Harvard Library project were mapped. The Working Group's analysis concluded that sufficient commonalities exist among the schemes to achieve consensus on a core set of preservation metadata. In June 2002, the Working Group published a framework for preservation metadata based on the common elements previously identified [OCLC/RLG, 2002]. An overview of the framework is presented in Appendix B.

Based on the work on preservation metadata, OCLC extended its CORC (Cooperative Online Resources Cataloging) system for cataloging Web resources to include preservation metadata. (The system has recently been renamed the OCLC Connexion [OCLC Connexion, 2002]). Catalogers at the U.S. Government Printing Office are being trained to use this system. As part of this project, team members saw a demonstration of the GPO system.

4.3.4 Technical Metadata

Technical metadata is a component of administrative metadata that helps to specify the format of the object. The metadata elements will vary depending on the format of the object – text, digital still images, audio, etc. The following discussion is not complete for all objects of interest to GSFC preservation activities, but instead is meant to serve as a example of the existing standards (or lack thereof) that will be encountered when making decisions about preservation formats. See the discussion of transformation versus native formats in Section 4.4.2.

4.3.4.1 Text Objects

Text objects may be received and stored in several formats. Most common are ASCII text, TIFF image and pdf. Each of these has special technical metadata associated with it that will be of importance for preservation and in particular for rendering the text in the future. For ASCII, it is important to know if the ASCII character set is the extended ASCII or perhaps Unicode. TIFF has several versions that should be identified in the technical metadata. Most common is TIFF IV. PDF files can vary by the version of Adobe Acrobat that was used to create them.

4.3.4.2 Digital Still Images

In addition there are relevant standards for specific object types. NISO and AIIM have developed a trial standard for digital still images [NISO, 2002]. This trial standard, which was issued in June 2002, documents the technical attributes of digital still images. Technical metadata is necessary to document image provenance and history (production metadata) and to ensure that image data will be rendered accurately on output whether to screen, print or film. It was also noted that preservation requires the development of applications to validate, process, refresh and migrate images against criteria that is based on the technical metadata.

The elements are divided into five basic groups. Basic image parameters, such as MIMEType, Compression, and Display Orientation, include information needed to reconstruct the digital file into a viewable image on an electronic display. Image Creation elements document the logistics and administrative conditions under which the digital image was captured. Elements in this set include SourceType, HostComputer, ScannerManufacturer, and DigitalCameraManufacturer. The Imaging performance assessment set provides attributes of images that are inherent to the quality. These elements serve as metrics to assess the accuracy of output and of preservation techniques, particularly migration. They include spatial metrics, colormap information, and target data. Change history documents the processes applied to the image over the life cycle. Processes result in either editing or transforming the image. The current information is not erased when adding new information to the image history. The data dictionary developed can be obtained from www.niso.org/standards/dsftu.html.

The trial standard is the basis for current work at the Library of Congress and at the Cornell University digital library projects.

4.3.4.3 Video

There are several video standards such as the Universal Preservation Format, originally developed by WGBH in Boston. Groups as diverse as Warner Brothers and the Department of Defense are working on standards related to video for entertainment, training and distance education. One of the major issues with regard to video is that there is no standard format for video. While MP4 is very common, it is also proprietary and there have been issues regarding the degree to which it is a good choice for preservation. Unfortunately, it is the one with the most tools available.

4.3.5 Permanence Ratings

As a contribution to the preservation metadata environment, the National Library of Medicine developed a permanence rating system to address a user's need to know whether a resource will remain available, unchanged and in the same location when needed in the future [NLM Working Group on Permanence, 2000]. The system is based on three indicators – identifier validity, resource availability, and content invariance (see Appendix C). The system allows content managers to convey this information concisely to humans or to a computer system accessing a resource. The Permanence Rating system from NLM is serving as a model for addressing retention issues. It was tested with a variety of NLM Web-based resources. The actual system at NLM is awaiting installation of software to support the archiving process. Other organizations are evaluating the use of the Permanence Ratings for their materials.

4.3.6 Other Applicable Standards

Indirect standards are those that have other purposes besides archiving but that are expected to serve important functions in the archiving infrastructure. Persistent identifiers, such as the Handle® [Coalition for National Research Initiatives, 2002], are needed to support reference linking and location of digital objects over time. The Open Archives Initiative, which was developed with e-print repositories in mind, provides a protocol for exposing consistent metadata and then harvesting a central metadata repository from remote compliant archives [Open Archives Initiative, 2002]. This may prove to be a significant component for interoperable digital archives, with the preservation metadata defined as an extension to the base OAI metadata specifically for the preservation community.

4.4 Technical Preservation Strategies

4.4.1 Migration and Emulation

There are three major methods for dealing with the ever-changing technology of the digital environment – technology preservation or computer museums, migration and emulation. Technology preservation maintains the computers, operating systems, copies of the application software, etc. required to provide access to specific digital objects. This is the equivalent of keeping a keypunch machine around to be able to access punched cards.

Migration is the most well established method. It involves moving digital content from one version of software to another and from one machine to another. However, as many have learned in the past, this approach can be less than satisfactory when the software or hardware is not of the commercial variety or where the organization has not kept up with the migration regime. Suddenly, the organization may find that migration is not supported because maintenance has not been paid or upgrades have not been installed.

Until recently, emulation has been discussed in theory [Rothenberg, 1999, 2000], but there were few practical attempts to determine its feasibility as a preservation method. However, emulation is based on concepts that have been used in mainframe environments for many years.

As an outgrowth of the Cedars Project, CAMiLEON (Creative Archiving at Michigan and Leeds: Emulating the Old on the New), a joint project of the University of Michigan and the University of Leeds, was established. The goal of the project was to determine the practical, long-term feasibility of emulation as an approach to preservation [CAMiLEON, 2001]. Using virtual machine technologies, this project has taken significant steps to prove that it is possible to run old software and its data on new machines.

However, CAMiLEON's major contribution may be the realization that both migration and emulation have their place as preservation strategies [Granger, 2000]. The appropriate preservation strategy may vary based on the content. In some cases migration is what is needed; in cases where there is complex interaction of the software, hardware and content, emulation may be the more appropriate approach. A more complete analysis of the uses of various strategies, including technology preservation, was done in order to outline costs for digital preservation [Hendley, 1998].

Based on the outcome of CAMiLEON, Cedars has suggested in its final guide to digital preservation strategies that a combined approach would reduce the cost of preservation and ensure the most important information is retained for later use [Cedars, 2002, April]. The Cedars approach preserves a bytestream with appropriate technical metadata. The technical metadata is consistent for a particular class of digital resource, so the effort of creating the technical metadata is amortized over all the resources in that class creating economies of scale. When the current version of the software changes, the technical metadata can be changed once, rather than changing the digital objects themselves. According to Cedars, this approach will eliminate the loss of information through successive migrations and reduce the risk recreating the technical environment via emulation will be unsuccessful.

In addition to guidelines, a physical system for archiving is being developed by IBM Netherlands. The Dutch National Library and the British Library are jointly funding this development. The system will be based on the OAIS Reference Model and incorporate many of the standards or quasi standards that have been developed through NEDLIB and other projects described above. While this effort is focused on electronic journals and other text materials, there will ultimately be a great deal of flexibility in the objects that can be handled. In addition, the system will be independent of the preservation approach – emulation or migration.

4.4.2 Transformation vs. Native Formats

A key preservation issue is the format in which the archival version should be stored. Transformation is the process of converting the native format to a standard format. On the whole, the projects reviewed favored storage in native formats. However, there are several examples of data transformation. American Astronomic Society, the National Library of Medicine and the American Chemical Society transform the incoming files into SGML or XML-tagged ASCII format. The AAS believes that "The electronic master copy, if done well, is able to serve as the robust electronic archival copy. Such a well-tagged copy can be updated periodically, at very little cost, to take advantage of advances in both technology and standards. The content remains unchanged, but the public electronic version can be updated to remain compatible with the advances in browsers and other access technology." [Boyce 1997]

The data community also provides some examples of data transformation. For example, the NASA Data Active Archive Centers (DAACs) transform incoming satellite and ground-monitoring information into standard Common Data Format. The UK's National Digital Archive of Datasets (NDAD) transforms the native format into one of its own devising, since it could not find an existing standard that dealt with all its metadata needs. These transformed formats are considered to be the archival versions, but the original copies are retained, so that someone can replicate what the center has done if necessary.

At the specific format level, there are several approaches used to save the "look and feel" of material. The majority of the projects reviewed use image files (TIFF), .pdf, or HTML for text. TIFF does not allow the embedded references to be active hyperlinks. For purely electronic documents, .pdf is the most prevalent format. This provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. While .pdf is increasingly accepted, concerns remain for long-term preservation and it may not be accepted as a legal depository format, because of its proprietary nature.

Many are considering XML as a scheme for preserving content since it allows for encoding of the content's meaning. However, in order to preserve the look and feel it must be properly developed to include preservation aspects needed to render the object in the future.

Cedars identified the aspects needed for reuse as ***significant properties*** [Cedars, 2002, April]. It is these significant properties that are important for future rendering to serve the needs of a particular community. In the case of text manuscripts, it may be that only the ASCII text is important. Therefore, all other aspects of the original may be stripped away. However, in other cases, the requirements of the formats or of potential use and users would dictate that the more aspects of the look and feel be preserved. For example, with an electronic journal, the links to outside references may be of particular importance.

Preserving the "look and feel" is difficult in the text environment, but it is even more difficult in the multimedia environment, where there is a tightly coupled interplay between software, hardware and content. The U.S. Department of Defense DITT Project is developing models and software for the management of multimedia objects. Similarly, the University of California at San Diego has developed a model for object-based archiving that allows various levels and types of metadata with distributed storage of various data types. The UCSD work is funded by the U.S. National Archives and Records Administration and the U.S. Patent and Trademark Office. These activities should be followed for their future applicability to GSFC's environment.

4.4.3 Authenticity and Validity

It is the responsibility of the Preservation Planning function, along with the Administration function of the OAIS Reference Model, to consider security and validation issues. How do we verify and ensure data integrity? How do we ensure completeness of the received data in electronic form? For example, there is concern among image archivists that images can be tampered with without detection. Clifford Lynch of the Coalition for Networked Information stated at a recent CENDI meeting (August 2002) that while redundancy may be desired in order

to ensure the continuity of archives, redundancy can also quickly propagate errors, whether intentional or unintentional. Particularly in cases where conservation issues are at stake, it is important to have metadata to manage encryption, watermarks, digital signatures, etc. that can survive despite changes in the format and media on which the digital item is stored.

On an almost global front, the International Research on Permanent Authentic Records in Electronic Systems (InterPARES), is a coalition of national, university and government agency archives [InterPARES, 2002]. There are several major groups involved in InterPARES, including regional members for Asia and Europe. The goal is to provide guidelines for the preservation of authentic electronic records, preserving the "place" of that object in the collection and ensuring its validity for legal purposes. Best practices, tools and standards specific to authenticity are being identified.

4.5 Organizational Models for Archiving

Early work by the NASA DAACs identified several models for archiving. It is worthwhile looking at the pros and cons for these and then later to analyze them with respect to the GSFC situation.

4.5.1 Centralized Repository

The centralized model spiders or harvests information from relevant sites and then copies the content of the sites into a central repository. Generally, the centralized repository takes control of the materials and provides long-term access mechanisms. The Internet Archive, sponsored by computer/information magnate Brewster Kahle, takes snapshots of the public surface Web. In October 2001, the Internet Archive enhanced access to the various snapshot collections through the WayBack Machine, search software that allows access by URL or by date range [Internet Archive, 2001]. This model is also available under a service agreement; the Library of Congress, for example, has received tapes from the Internet archive for preservation purposes. The Internet Archive collects the results of the spidering centrally and then turns the repository over to the local institution for preservation.

4.5.2 Third-Party Repositories

Third-party repositories are a special kind of centralized model. These organizations are service organizations that are not themselves involved in digital creation or publishing. OCLC's Digital Archive is such a trusted third-party repository [OCLC Digital Archive, 2002]. OCLC provides the tools for archiving, storing, accessing and, in some cases, handling the licensing or copyright issues that may be relevant to a particular environment for the long-term. The current cost model for the Digital Archive (and the OCLC Connexion which provides underlying metadata input services) is based on a cost per record added and an annual fee for storage of the digital objects.

In addition, OCLC recently announced the Digital & Preservation Co-op [OCLC Digital Preservation Resources, 2002]. The goal of the Co-op is to build collections and knowledge through collaboration and to save money or to find grant opportunities for members through Co-op participation. The Co-op will deal with both digitized and "born digital" materials. The Co-

op's business model involves a membership fee on the part of the producing (or publishing) organization. While the fees have been waived for charter members, the regular cost will be approximately $1000. The GSFC Library is a charter member of this group and the Library Director recently attended the initial meeting of the group in Dublin, OH.

4.5.3 Federated Repositories

The federated model is the opposite of the centralized model. It calls for a distributed environment made up of a number of archive nodes that are linked together by formal or informal agreements, but minimally through some standards that provide interoperability. The NASA DAACs are examples of a federated approach. (See the previous discussion of the DAACs in Section 3.3.2 for a description of how the centers are federated.) Through minimal standards, common search systems, and common data management tools, the repositories can interoperate while retaining their independence and ability to customize for their own particular clients. This federation also allows for redundancy in case of disaster.

5.0 RESULTS OF THE GODDARD LIBRARY PILOT PROJECTS

As part of the Digital Archiving Project, two pilot projects were conducted. The first captured video from GSFC colloquia, lectures and mini-courses. The second captured GSFC Web sites, with a particular focus on Web sites for projects.

5.1 Video Capturing

The purpose of the pilot project was to capture the content of these colloquia and treat this content as a GSFC knowledge asset. The goal was to catalog, index, store, preserve and provide access to the content of these colloquia to the GSFC audience from the desktop. This pilot involved handling the existing content stored on videotapes, as well as the "born digital" assets created through video streaming. The details of this project are provided in Appendix D.

In order to effectively capture the content created in the colloquia series a concerted effort had to be devoted to communicating with all of the managers of the colloquia series to promote the benefits of distributing Web content. All the managers were contacted in order to discuss how the library could provide access to the content created through each series, and to offer the Library's services for organizing, delivering and archiving.

The methodology for handling new colloquia series was piloted with the IS&T colloquia series. The live analog broadcast is converted into a digital file by WindowsMedia 7.1 encoder. The encoder adds the following fixed set of metadata elements to the file during this process: Author, Title, Copyright, and Abstract. The digital file is then sent to a MediaMan server in the library running Windows 2000 for streaming using Windows IIS Webserver and Windows Media Administrator. This server is used to provide better streaming performance. By using this server to serve the content to the desktop it reduces the load on the encoding computer. When the file on the encoder is complete it is ftp'd to another computer for editing, cleanup and archiving. An .asx file is created to provide persistent urls for these assets. The .asx file contains any scripts that might be associated with the file and the metadata for the file. All stored content is archived

on the MediaMan server. Access to the stored files is controlled through IIS and falls into 3 categories (directories): Goddard only, NASA only, public. This is controlled through an IP check.

Each colloquium is videotaped and a copy is sent to the Library for cataloging. The tapes are assigned bibliographic descriptions and put into a database that is accessible via the GSFC Web. As part of this pilot consideration was given to improving the access to the content of these tapes by transferring them to digital format and delivering them through the same Web interface as the other digital ones. The methodology for this conversion utilized the same software products used for the live broadcasts. Videotapes were run through a standard VCR via s-video to the encoder. This has to be done in real-time.

During FY2001 the GSFC Library began live Webcasts of the IS&T Colloquia to the internal GSFC audience. The process of transforming the taped broadcast into a digital file and providing access to the stored content of that file became a way to capture the content of these presentations, including the question and answer sessions following. These GSFC-created knowledge assets are available via the Web for future use.

At the beginning of the project the Library was involved in only one of the colloquia series (IS&T.) As a result of having a staff person assigned to coordinating and promoting the advantages of digital content being provided via the Web, the Library now has some involvement in all of the colloquia series. The new Systems Engineering Seminar series have both live Webcasts and stored content for all of their presentations. The Library is providing that service to this series.

5.1.1 Quality of Original Video Output

One of the first lessons learned was related to the video creation of the assets. The Technical Services Branch manages all segments of the creation of these assets. The camera angles, audio feeds and lighting are outside of the control of the Library. This can create some issues when the quality of the original source materials has been compromised.

5.1.2 Video Feed Speed

Initially the bit-rate that was being used for the digital creation was monobit (one stream) at 218KBS and 320x240 pixels. This rate produced a high-speed feed but was not acceptable for a slower dial-up connection. As a result of that experience a second copy had to be produced to support the lower streaming rate of 15KBS. An improved process of using a multi-bit stream, which allowed for multiple streams at multiple rates, was implemented.

5.1.3 File Sizes

The files created by this process are very large. Backup of the media server where these are stored is not feasible without very large disk arrays. Alternative storage plans must be developed to support long-term preservation.

5.1.4 Indexing

As the process progressed it became clear that indexing the content of the presentation was not a simple task. Technologies that are available to convert speech to text without human intervention are not mature enough to provide this capability. As a result of the pilot project software was purchased for testing and application to the future colloquia.

Speech to text recognition became a major part of this project toward the end. This capability is needed to provide content that can be searched via text-based search engines. One way around this has been the development through MS products of a way to key the PPT slides to the Webcast. This allows the user to not only see the visuals at the same time that he/she hears the speaker talk about them, but it also provides some text from the PPT slides that can be used as text for searching.

The Inmagic product used for this pilot has minimal metadata elements available. We tried to extend the metadata to include more elements needed for long-term preservation. This has not proven successful to date. However, there may be some ways to subset the existing fields in order to provide additional information in an otherwise stable record. Alternatively, it may be possible to link the metadata record as an external file to the brief metadata record in Inmagic, which is essentially geared to discovery and retrieval, rather than preservation.

5.1.5 Providing Access to Slide Presentations

In addition, it is rather difficult to understand the lecture when the speaker relies heavily on printed materials and the camera is focused on the person. Providing a split screen, which includes both the speaker and his/her PPT presentation, is in the works.

5.1.6 Lack of Video Format Standards

Long term archiving of video content will probably involve many migrations from format to format. There is concern that with each migration a degradation of quality may occur. In the creation of the archive file consideration should be made of what quality/resolution should be used for the original versus the quality of the file streamed to the user. At present the highest quality possible from the current system is 740x620. A test should be run to see if it is possible to use the high resolution to create two multi-bit streams to accommodate the bandwidth available for delivering the content.

Long-term preservation needs to be a part of the video capture program. A plan is in place to create two dvd copies of each digital asset. One copy will be stored in the Library to provide as backup, and the second will be stored at the off-site storage facility.

5.2 Web Page Capturing

A pilot project was also conducted on the archiving of Web pages. The details of this pilot project are provided in Appendix E. Project Web pages will become increasingly important as

results are published to the Web and as groups track, perform and develop their projects via project Web sites, groupware applications and collaboratoria.

The team conducted tests on several types of sites including GSFC library pages, GSFC project pages selected from the GSFC Project Directory, and non-project pages consisting of pages from the scientific codes. The library pages (in the library.gsfc.nasa.gov domain) were used primarily as test material to help in establishing the setting that should be used for the capture of project and scientific Web sites. The Library site has little original material that it would want to provide to the public.

The MAP Web site (http://map.gsfc.nasa.gov/) is a NASA Explorer Mission that will measure the temperature of the cosmic background radiation over the full sky with unprecedented accuracy. RHESSI's (http://hesperia.gsfc.nasa.gov/hessi/) primary mission is to explore the basic physics of particle acceleration and explosive energy release in solar flares. The Technology page (http://gsfctechnology.gsfc.nasa.gov/) is a science page available from one of the directorates. The content includes: Current NASA activities, Technology investment areas, Distributed Space Systems, Flight & Science Information Systems, etc. The site is geared toward explaining some of the more gritty technology used at NASA to the public so they can understand what NASA is doing and how. The TDRS program site (http://nmsp.gsfc.nasa.gov/tdrss/tdrshij.html/) was used initially but caused some problems with links to outside resources, so it was deleted from the final analysis. The Tracking and Data Relay Satellites comprise the space segment of NASA's communications relay system, providing telecommunication services to low earth orbiting spacecraft.

The major lessons learned are highlighted below.

5.2.1 Spidering or Crawling Web Sites

The team investigated various types of spidering software. However, because of limited resources, the team decided to use the freeware product, HTTracker. Other similar programs were identified, but an initial analysis indicated that for the cost of the software there was little functionality to be gained. The URLs for the project homepages were provided to the HTTracker software as starting points for its spider. It then copies and indexes any site that are publicly available using settings, which can either be set to the default or customized. The major settings were to not accept cookies, to obey robot,txt rules for no access, the level from start page from which spidering should occur, etc. The team then analyzed the results and the log information to identify issues and problems.

5.2.2 Page Access Problems

Page Access Problems are indicated by two errors – 404 and 403. The first are hyperlinks that are broken. This can result from the pages having been deleted or moved to another physical location such as another server, the directory structure or new file name. In testing the project directory sites, the team found 11 broken links. Upon further investigation, we found that more than half of them had moved and we were able to find them. However, the type of investigative work needed to locate the pages is prohibitive in an operational system. This is why support

from the project librarians and the institution of a persistent identifier system is needed (see Section 4.3.6).

403 errors result when the linked pages are at the location but the spider cannot access them. These errors may result from portions of the site that have been blocked to robots and spiders, etc. (The spider has the ability to bypass these robot prohibitions, but such practice is not recommended.) Also, the sites with 403 errors may be password protected or require a specific IP address to gain access. One 403 error was found in the sample of project and non-project sites. There were several 403 errors from the Library site, because many library resources from external sources require passwords.

5.2.3 Extent of the Sites

To what level should the spiders crawl? This is an issue because of the wide use of links. The number of pages crawled, copied and mirrored grows exponentially with each level that is followed. Some guidance can be found in other projects. In this case, decisions could be made based on a farm of servers that belong to GSFC, to NASA, or to project partners.

Intellectual property issues can occur even within the GSFC domain because of the number of non-government partners with which GSFC interacts. While work has been done to develop standards for metadata related to intellectual property rights (<indecs>), the implementation of a system that moderates these rights can be costly. It may be worthwhile in follow up work to analyze solutions that might involve the way that grants and contracts are written.

The spider ran out of total time to perform the session. This was fixed by not setting a limit, but this means that resources are being used somewhat uncontrollably.

For the three project sites that were scanned to the fifth level, over 1.5 gigabytes were captured. The scan took more than a day and a half and was not finished when it was cancelled. One of the sites had scanned only 5,333 links of a total of over 40,000 links after a day and a half.

The team found a variety of information at project sites, particularly from the links to the pages. Should video and non-text files be scanned and incorporated in the same way? Many of these non-text files are extremely large, but they also appear to be important content. For example, one of the movie files grabbed by the spider was more than 174 MB in size.

On the whole, the size of the files resulting from the spidering may be prohibitive. Despite the fact that storage is relatively cheap, the ongoing management of files of this size requires that some further analysis be done of how the spidering could be more customized.

This raises issues about the spidering software that was used. While HTTrack appeared to be the best for the test purposes, it does not provide sufficient statistics and control of the types of sites (.com, .org, etc.) that are grabbed at lower levels to provide the controlled crawling that would be needed to perform this kind of customization and to efficiently use the GSFC resources.

5.2.4 Deep Web Content

While this pilot project only dealt with Web pages, many of these pages contain content that is accessible via the Web but not directly. This is the so-called hidden or deep Web, represented by a variety of document types that are not HTML. The deep web includes databases, pdf files, and software programs. While some of these object types would be handled through archiving mechanisms related to other types of content, a traditional spidering method is unlikely to adequately handle these object types. The most problematic of these sites require proprietary software to use. This severely impacts the future access to this information and requires special procedures, including perhaps the development of a software registry that could be used to link objects of these types to the software that should be loaded in order to use them.

5.2.5 Dynamic Web Pages

Generally, spiders are only able to deal with pages that physically exist. Active server pages and pages that are created on-the-fly from content management and portal management systems are dynamic. There is no physical content to grab and copy. While the sample of the GSFC Web sites did not include any dynamic page generation, the use of dynamic page techniques are likely to increase as interest in and use of portals and content management increases in the future.

5.2.6 Metadata Creation

To a lesser extent than with video capture, there are issues related to metadata creation. The pilot project used dc.dot software from UKOLN to automatically create a template of Dublin Core elements from the HTML content of the page. In addition, the dc.dot software provides a mechanism for converting the dc.dot HTML output to XML for import into a structured database. This automatic approach creates metadata elements only to the degree that the page is well formed in the first place. The analysis of the sample project Web pages showed that approximately 20% of the functioning sites surveyed have HTML metatags. Examples of dc.dot output and XML conversion are provided in Appendix G.

5.2.7 System Set Up

The set up for the pilot project was a networked machine used for other purposes. Therefore, many of the processes had to be done at night when there was no one in attendance to de-bug or stop the spider when it was "going haywire". The bandwidth of the Internet connection was adequate but the speed of the machine was not. In the latter part of the project a replacement machine was procured with a higher speed processor and more storage. This significantly improved the spidering.

The team recommends an isolated machine that is dedicated to this process. In addition, it should have significant processor speed, storage available for processing and caching, and external storage to accommodate the archived content.

5.2.8 Virus Protection

During the process, the team encountered a problem with a commercial company that was being spidered by the process. It took a long time to manually analyze the logs created by HTTrack to determine how and why this occurred. It still isn't clear whether there was a virus on the machine or if someone spoofed the IP from the system and used it to try to hack into the commercial company's computer.

In any case, viruses can be a significant problem, because any copy of information from another computer can include a virus. In an operational system, virus protection software should be installed and every file should be checked before being included in the actual digital archiving system. This is another reason for having an isolated computer as the initial capture point.

5.2.9 Retrieval and User Interface Design

As part of the project, the team took the results of the spidering, indexed it via Autonomy's Server search engine, and presented it via a user interface. For purposes of initial testing, the default values of Autonomy were used, and all files were indexed, including the pages from the spidering of the Library site and the TDRS site. The standard Autonomy Server interface was used without modification.

Based on this initial testing, the team identified the following issues:

5.2.9.1 Autonomy Index Creation Time

The size of the files anticipated for a digital archiving system will, if indexed in total, present a major indexing effort under Autonomy. This is not only a result of Autonomy but of the system on which the Library is running Autonomy.

5.2.9.2 Non-Text Files

If the native content from the spidering is provided to the Autonomy engine, it will include many movie and graphics files that are not appropriate for Autonomy text-based searching. No metadata has been applied to these objects, and the resources are not available to do this manually. A mechanism must be devised to remove these files prior to indexing but to ensure that they are retained in the archiving system and available when the user accesses the archive as the result of a search. If the creator or originating system creates metadata, it should be made available to Autonomy's indexing process.

5.2.9.3 Text Files in Proprietary Formats

The problem also exists when dealing with text files. For example, for each pdf file, Autonomy must first convert the pdf to binary and then perform the indexing. This is an extremely time consuming effort. The pdf files should be analyzed to determine the importance of having the text content of these files indexed and available for retrieval.

If the analysis determines that these files are not important for retrieval, then a process should be developed to ensure that they are not indexed. These are the same kinds of files that have issues related to long-term preservation of native versus transformed formats.

5.2.10 Retrieval of Objects Versus Collections

The Library learned in the CIO Pilot Project that users are interested in not only searching information across projects, but in grouping information by the project name or mission. A major issue that arises when developing the user interface for the archive is the degree to which the interface provides access by project collection versus individual object with a collection. For example, an interface could be envisioned that groups the project collections and then allows the user to browse through the retained information under that project and to search within only that collection. The METS collection level metadata design would support the structure of such an archive. However, there is also the issue of identifying the collection of interest in the first place and satisfying the needs of users who want topical information regardless of the project involved.

6.0 CONCLUSIONS AND NEXT STEPS

After collection of the information from within and outside Goddard as well as the analysis of the results from the two pilot projects, the group identified the following next steps.

- Develop a framework document for digital archiving at GSFC (see Part 2 of this report) which considered issues raised by the benchmarks of other external systems and the findings of the two pilot projects
- Develop a staged implementation plan
- Develop a model similar to that used by the National Agricultural Library to link standards in the GSFC environment under the OAIS RM framework
- Identify how the results of this project interrelate with other activities within GSFC
- Perform a gap analysis to determine a research agenda

7.0 REFERENCES

Boyce, P. (1997, November). Costs, Archiving and the Publishing Process in Electronic STM Journals. Against the Grain, 9(5), p. 86. Retrieved May 3, 2002 from the American Astronomical Society Web site: www.aas.org/~pboyce/epubs/atg98a-2.html

CAMiLEON: Creating Creative Archiving at Michigan & Leeds: Emulating the Old on the New. (2001). Retrieved May 3, 2002 from the CAMiLEON Web site: www.si.umich.edu/CAMILEON/

Cedars: CURL Exemplars in Digital Archives. Retrieved May 3, 2002 from the University of Leeds Web site: www.leeds.ac.uk/cedars/

Cedars. (2002, April). Cedars Guide to: Digital Preservation Strategies. Retrieved September 29, 2002 from the Cedars Web site: http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html

CENDI. (2001). Managing and Preserving Electronic Resources: The OAIS Reference Model. A symposium jointly sponsored by CENDI and the Federal Library and Information Center Committee, December 11, 2001. Retrieved September 29, 2002 from the CENDI Web site: http://www.dtic.mil/cendi/activities/12_11_01_oais_program.html

Coalition for National Research Initiatives. (2002). Handle System. Retrieved September 30, 2002 from the CNRI Web site: www.handle.net/

Committee on an Information Technology Strategy for the Library of Congress. Computer Sciences and Telecommunications Board, National Research Council. (2001). LC21: A Digital Strategy for the Library of Congress. National Academy Press: Washington DC. Retrieved May 3, 2002 from the National Academy Press Web site: books.nap.edu/books/0309071445/html/index.html

Consultative Committee for Space Data Systems (CCSDS). (2002). Reference Model for an Open Archival Information System (OAIS). Retrieved September 29, 2002 from the ISO Archiving standards/NASA Goddard Web site: ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

Goddard Space Flight Center Library Visiting Committee Report. (2002).

Granger, S. (2000). Emulation as a Digital Preservation Strategy. D-Lib Magazine, 6(10). Retrieved May 3, 2002 from the D-Lib Magazine Web site: www.dlib.org/dlib/october00/granger/10granger.html

Hendley, T. (1998) Comparison of Methods and Costs of Digital Preservation. Prepared for the Joint Information Systems Committee and the British Library Research and Innovation Centre. Retrieved August 26, 2002 from the UKOLN Web site: www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html

Inera, Inc., (2001). E-Journal Archive DTD Feasibility Study. Prepared for the Harvard University Library, Office of Information Systems E-Journal Archiving Project. Retrieved May 3, 2002 from the Digital Library Web site: www.diglib.org/preserve/hadtdfs.pdf

Internet Archive: building an 'Internet Library'. (2001). Retrieved May 3, 2002 from the Internet Archive Web site: www.archive.org

InterPARES: International Research on Permanent Authentic Records in Electronic Systems. (2002). Retrieved May 3, 2002 from the InterPARES Web site: www.interpares.org

Jones, M. & N. Beagrie. (2001). Preservation Management of Digital Materials: A Handbook. London: The British Library.

Lounamaa, K. and I. Salonharju. (1999, January). "EVA-The Acquisition and Archiving of Electronic Network Publications in Finalnd." Tietolinja News, 1. Retrieved May 3, 2002 from the University of Helsinki Web site: www.lib.helsinki.fi/tietolinja/0199/evaart.html

National Agricultural Library Metadata Task Force. (2001). Template for Standard NAL Metadata. Retrieved September 29, 2002 from NAL Web site: http://www.nal.usda.gov/cataloging/TEMPLATE2.pdf

National Information Standards Organization. (2002, June). Technical Metadata for Digital Still Images. (Trial draft). Retrieved September 17, 2002 from the NISO Web site: www.niso.org/standards/dsftu.html

National Library of Medicine Working Group on Permanence of NLM Electronic Publications. (2002, October). Phase II Report. Retrieved September 29, 2002 from the National Library of Medicine Web site: www.nlm.nih.gov/pubs/reports/permanence.pdf

Networked European Deposit Library: NEDLIB. (2001). Retrieved May 3, 2002 from the Koninklijke Bibliotheek Web site: www.konbib.nl/nedlib/

OCLC Connexion. (2002). Retrieved September 29, 2002 from the OCLC Web site: www.oclc.org/connexion/

OCLC Digital Archive. (2002). Retrieved May 3, 2002 from the OCLC Web site: www.oclc.org/digitalpreservation/about/archive/

OCLC Digital Preservation Resources, Digital & Preservation Co-op. (2002). Retrieved May 3, 2002 from the OCLC Web site: www.oclc.org/digitalpreservation/about/co-op/

OCLC/RLG Working Group on Preservation Metadata. (June, 2002). A Framework….. Retrieved September 29, 2002 from the OCLC Web site: http://www.oclc.org/research/pmwg/pm_framework.pdf

Open Archives Initiative. (2002). Retrieved September 25, 2002 from the OAI Web site: www.openarchives.org

PANDORA. Retrieved May 3, 2002 from the National Library of Australia Web site: pandora.nla.gov.au/index.html

Research Libraries Group. (2001, August). Attributes of a Trusted digital Repository for Digital Materials: Meeting the Needs for Research Resources. Retrieved May 3, 2002 from the Research Libraries Group Web site: www.rlg.org/longterm/attributes01.pdf

Rothenberg, J. (1999, January). Avoiding Technological Quicksand: finding a Viable Technical Foundation for digital Preservation. Report to CLIR. Retrieved May 3, 2002 from Council on Library and Information Resources Web site: www.clir.org/pubs/reports/rothenberg/contents.html

Rothenberg, J. (2000, April). An Experiment in Using Emulation to Preserve Digital Publications. NEDLIB Report Series; 1. Retrieved May 3, 2002 from the NEDLIB Web site: www.kb.nl/coop/nedlib/results/NEDLIBemulation.pdf

Royal Library. National Library of Sweden. (n.d.) Kulturaw3 – Heritage Project: Long Term Preservation of Published Electronic Documents. Retrieved May 3, 2002 from the National Library of Sweden Web site: www.kb.se/END/kbstart.htm

APPENDIX A
Dublin Core Elements, Version 1.1

Element: Title

Name: Title
Identifier: Title
Definition: A name given to the resource.
Comment: Typically, a Title will be a name by which the resource is formally known.

Element: Creator

Name: Creator
Identifier: Creator
Definition: An entity primarily responsible for making the content of the resource.
Comment: Examples of a Creator include a person, an organisation, or a service.
Typically, the name of a Creator should be used to indicate the entity.

Element: Subject

Name: Subject and Keywords
Identifier: Subject
Definition: The topic of the content of the resource.
Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource.
Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

Element: Description

Name: Description
Identifier: Description
Definition: An account of the content of the resource.
Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

Element: Publisher

Name: Publisher
Identifier: Publisher
Definition: An entity responsible for making the resource available
Comment: Examples of a Publisher include a person, an organisation, or a service.
Typically, the name of a Publisher should be used to indicate the entity.

Element: Contributor

Name: Contributor
Identifier: Contributor
Definition: An entity responsible for making contributions to the content of the resource.
Comment: Examples of a Contributor include a person, an organisation, or a service.
Typically, the name of a Contributor should be used to indicate the entity.

Element: Date

Name: Date
Identifier: Date
Definition: A date associated with an event in the life cycle of the resource.
Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.

Element: Type

Name: Resource Type
Identifier: Type
Definition: The nature or genre of the content of the resource.
Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types [DCT1]). To describe the physical or digital manifestation of the resource, use the FORMAT element.

Element: Format

Name: Format
Identifier: Format
Definition: The physical or digital manifestation of the resource.
Comment: Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).

Element: Identifier

Name: Resource Identifier
Identifier: Identifier
Definition: An unambiguous reference to the resource within a given context.
Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system.

Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).

Element: Source

Name: Source
Identifier: Source
Definition: A Reference to a resource from which the present resource is derived.
Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Element: Language

Name: Language
Identifier: Language
Definition: A language of the intellectual content of the resource.
Comment: Recommended best practice for the values of the Language element is defined by RFC 1766 [RFC1766] which includes a two-letter Language Code (taken from the ISO 639 standard [ISO639]), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard [ISO3166]). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.

Element: Relation

Name: Relation
Identifier: Relation
Definition: A reference to a related resource.
Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Element: Coverage

Name: Coverage
Identifier: Coverage
Definition: The extent or scope of the content of the resource.
Comment: Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.

Element: Rights

```
Name:       Rights Management
Identifier: Rights
Definition: Information about rights held in and over the resource.
Comment:    Typically, a Rights element will contain a rights
            management statement for the resource, or reference
            a service providing such information. Rights information
            often encompasses Intellectual Property Rights (IPR),
            Copyright, and various Property Rights.
            If the Rights element is absent, no assumptions can be made
            about the status of these and other rights with respect to
            the resource.
```

APPENDIX B
Preservation Metadata Elements

The following is a list of the preservation metadata framework recommended by OCT/RLG Working Group on Preservation Metadata. Full descriptions for each section of the framework and definitions and examples for individual elements can be found in the main body of the framework report at http://www.oclc.org/research/pmwg/pm_framework.pdf.

The high-level framework is shown in ALL CAPS FONT. Metadata elements are in bold and sub-elements are in regular font.

CONTENT INFORMATION

CONTENT DATA OBJECT

REPRESENTATION INFORMATION
CONTENT DATA OBJECT DESCRIPTION
Underlying abstract form description
Structural type
Technical infrastructure of complex object
File description
Installation requirements
Size
Access inhibitors
Access facilitators
Significant properties
Functionality
Description of rendered content
Quirks
Documentation
ENVIRONMENT DESCRIPTION
SOFTWARE ENVIRONMENT
RENDERING PROGRAMS
Transformation process
Transformer engine
Parameters
Input format
Output format
Location
Documentation
Display/access application
Input format
Output format
Location
Documentation
OPERATING SYSTEM
OS name
OS version
Location
Documentation
HARDWARE ENVIRONMENT
Location
COMPUTATIONAL RESOURCES
Microprocessor requirements

Memory requirements
Documentation

STORAGE
Storage information
Documentation

PERIPHERALS
Peripheral requirements
Documentation

PRESERVATION DESCRIPTION INFORMATION

REFERENCE INFORMATION
Archival system identification
Value
Construction method
Responsible agency
Global identification
Value
Construction method
Responsible agency
Resource description
Existing metadata
Existing records

CONTEXT INFORMATION
Reason for creation
Relationships
Manifestation
Relationship type
Identification
Intellectual content
Relationship type
Identification

PROVENANCE INFORMATION
Origin
Event
Designation
Procedure
Date
Responsible agency
Outcome
Note
Next occurrence
Pre-ingest
Event
Designation
Procedure
Date
Responsible agency
Outcome
Note
Next occurrence
Ingest
Event
Designation

Procedure
Date
Responsible agency
Outcome
Note
Next occurrence

Archival retention

Event

Designation
Procedure
Date
Responsible agency
Outcome
Note
Next occurrence

Rights management

Event

Designation
Procedure
Date
Responsible agency
Outcome
Note
Next occurrence

FIXITY

Object Authentication

Authentication type
Authentication procedure
Authentication date
Authentication result

APPENDIX C
NLM PERMANENCE RATING SYSTEM

The proposed NLM permanence rating system includes three core categories for electronic resources: identifier validity, resource availability and content invariance. The definitions are available from the Appendix A of the Working Groups Phase II Report (http://www.nlm.nih.gov/pubs/reports/permanence.pdf). Each resource would contain information for all three, based on the controlled domain content provided under each core category.

IV: Identifier Validity
1. Transient
2. Guaranteed

RA: Resource Availability
1. No guarantee
2. Permanently available

CI: Content Invariance
1. Dynamic
   a. Growing
   b. Closed
2. Stable
   a. Growing
   b. Closed
3. Unchanging

APPENDIX D
Report On Video Capture Pilot Project

Purpose and Background:
GSFC has an extensive list of colloquia presented on center each year. These topics range from management issues to highly technical discussions. The purpose of the pilot project was to capture the content of these colloquia and treat this content as a GSFC knowledge asset. The goal was to catalog, index, store, preserve and provide access to the content of these colloquia to the GSFC audience from the desktop. This pilot should provide valuable experience with the issues surrounding digital video products as part of the overarching GSFC Digital Archive Plan and the GSFC Knowledge Management program

Methodology:
This project involved multiple approaches to addressing this complex issue. Each part of the process required a specific approach which in many ways evolved as the pilot progressed. This pilot involved handling the existing content stored on videotapes, as well as the "born digital" assets created through video streaming.

A. Cultural Issues:

GSFC is a very decentralized organization. Each small group maintains their our Websites, databases, etc. and there is limited involvement center wide. Additionally the scientific culture tends to promote independent development and a lack of sharing across organizations or projects. In order to effectively capture the content created in the colloquia series a concerted effort had to be devoted to communicating with all of the managers of the colloquia series and promoting the benefits of distributing Web content.

The methodology for this approach involved contacting all of the managers and discussing how the library can provide access to the content created through each series. The Library offered their services in organizing, delivering and archiving these assets.

B. Process for new colloquia:

The methodology for handling new colloquia series was piloted with the IS&T colloquia series. Library staff cooperated with GSFC Technical Services Bureau (TSB-media services) to get a live analog broadcast of the colloquia. This live feed is converted into a digital file using an s-video hookup and an Osprey 220 video card by Windows Media 7.1 encoder. The encoder adds a fixed set of metadata to the file during this process. These fields are:

Author
Title
Copyright
Abstract

The digital file is then sent to a MediaMan server in the library running Windows 2000 for streaming using Windows IIS Webserver and Windows Media Administrator. This server is used to provide better streaming performance. By using this server to serve the content to the desktop it reduces the load on the encoding computer.

When the file on the encoder is complete it is ftp'd to another computer for editing, cleanup and archive. An .asx file is created to provide persistent urls for these assets. The .asx file contains any scripts that might be associated with the file and the metadata for the file.

All stored content is archived on the MediaMan server. Access to the stored files is controlled through IIS and falls into 3 categories (directories): Goddard only, NASA only, public. This is controlled through an IP check.

C. Existing videotapes:

Each colloquia is videotaped and a copy of the video is sent to the GSFC Library for cataloging. The tapes are assigned bibliographic descriptions and put into a database that is accessible via the GSFC Web. As part of this pilot consideration was given to improving the access to the content of these tapes by transferring them to digital format and delivering them through the same Web interface as the other digital ones.

The methodology for this conversion utilized the same software products used for the live broadcasts. Videotapes were run through a standard VCR via s-video to the encoder. This has to be done in real-time.

Results:
This pilot would be termed a success. During FY2001 the GSFC Library began live Webcasts of the IS&T Colloquia to the internal GSFC audience. The process of transforming the taped broadcast into a digital file and providing access to the stored content of that file became a way to capture the content of these presentations, including the question and answer sessions following. These GSFC-created knowledge assets are available via the Web for future use.

At the beginning of the project the Library was involved in only one of the colloquia series (IS&T.) As a result of having a staff person assigned to coordinating and promoting the advantages of digital content being provided via the Web the Library now has some involvement in all of the colloquia series. The Library is also cooperating with several other GSFC organizations as a result of this project involvement. GSFC staff now has access to stored video content from all of the colloquia Websites for some of the FY02 speakers.

The new Systems Engineering Seminar series have both live Webcasts and stored content for all of their presentations. The Library is providing that service to this series.

As a side development of this pilot the Library has become more active in the Goddard Knowledge Management program. The pilot was investigating methods of access for the content of the video. The Goddard Knowledge Management Officer was also interested in this and so a partnership developed. This included exploring automatic indexing of the video content. Several software vendors were invited to demo their products for video indexing and speech-to-text conversion.

Lessons Learned:
One of the first lessons learned was related to the video creation of the assets. The TSB manages all segments of the creation of these assets. The camera angles, audio feeds and lighting are outside of the control of the Library. This can create some issues when the quality of the original source materials is inferior.

Initially the bit-rate that was being used for the digital creation was monobit (one stream) at 218KBS and 320x240 pixels. This rate produced a high speed feed but was not acceptable for a slower dial-up connection. As a result of that experience a second copy had to be produced to support the lower streaming rate of 15KBS. An improved process of using a multi-bit stream allowed for multiple streams at multiple rates was implemented and is now being used.

The files created by this process are very large. Backup of the media server where these are stored is not feasible without very large disk arrays. Alternative storage plans must be developed to support long-term preservation.

As the process progressed it became clear that indexing the content of the presentation was not a simple task. Technologies that are available to convert speech to text without human intervention are not mature enough to provide this capability. As a result of the pilot project software has been purchased to be piloted in FY03 for application to the future colloquia.

It is rather difficult to see the printed materials when the speaker relies heavily on printed materials and the camera is focused on them. (See next steps.)

Next Steps:
Long-term preservation needs to be a part of the video capture program. A plan is in place to create 2 dvd copies of each digital asset. The live copy will be delivered from the server as it is now. One dvd copy will be stored in the Library to provide as backup. The second dvd copy will be stored off-center, at the off-site storage facility.

A software product called StreamSage has been purchased for use in translating speech to text for improved indexing of the video assets. This product is supposed to allow for direct access to the frame or section of the file

that relates to the content being searched. Additional studies will be performed to determine if this is applicable for wide implementation.

Providing a split screen which includes the speaker and their powerpoint presentation is in the works. This depends on the cooperation of the TSB technicians, Oden networking support, and the speaker.

Several colloquia have allowed for sharing stored content on the Web but have not allowed live streaming. The Center Director's Colloquia is one that is planning on allowing live broadcast to the remote locations of Wallops and IV&V. The Library will not be directly involved in this live streaming but through the efforts of this pilot greater visibility for Webcasting has been promoted as a benefit to the center.

Long term archiving of video content will probably involve many migrations from format to format. There is concern that with each migration a degradation of quality may occur. In the creation of the archive file consideration should be made of what quality/resolution should be used for the original vs the quality of the file streamed to the user. At present the highest quality possible from the current system is 740x620. A test should be run to see if it is possible to use the high resolution to create two multibit streams to accommodate the bandwidth available for delivering the content.

APPENDIX E
Report On Web Capturing Pilot Project

Methodology:
The Web Capture Pilot included 4 subprojects or analyses. These included:

- Survey of the project Web sites
- Spidering sample project and science sites and collecting statistics
- Indexing the spidered contents and developing a metadata framework
- Developing a pilot semi-automated production environment for Web capture

Survey of the project Web sites

This analysis began by reviewing the first and fifth projects for each year in the Projects Directory developed and maintained by the GSFC Library. If there were not five projects for any given year, the last project of the year was used. The survey (see Appendix F) collected the following information: Link types, whether or not the project site contained any HTML metatags, types of multimedia contained on the site, the originating GSFC directorate code, and any extra comments about the site. This analysis provided key information about the types and formats that would need to be preserved and for which metadata elements and archival best practices should be surveyed or developed.

Of the 60 sites surveyed, only 46 of them were functioning. The years that recorded errors were 1997, 1995, 1993, 1992, 1989, 1988, 1983, 1978,1977, 1975(x2) and 1974. All of these errors were 404's. After 1995, there were only three errors, two 404's and one 403. Before 1995, there were 9 404's, which is also to be expected due to the age of some of the Web sites that are previous 1995, for they are either no longer in use or have been moved over the years.

Only two of the 60 project Web sites surveyed contained an originating GSFC directorate/code that could be found on the home page.

Of the 46 functioning sites, 42 were public friendly. Public-friendly was defined as information that was easy to understand and actually accessible to the public, or the provision of information that made the Web site easier to understand, such as additional links, a history of the project, reasons for the project and in some cases links for teachers. There were five pages that seemed to be geared only towards those concerned with the projects, dealing with complicated data sets and heavy technical jargon.

About half of the functional project sites surveyed (21) were front pages with links to other pages and/or sites. Nearly all of the front pages contained some sort of vital information dealing with the project itself.

All of the functional project sites surveyed contained some sort of link. Most links were to other pages within the project site. However, there were many different kinds of links, ranging from image galleries to data archives, to education/outreach, to charts and graphs. **Most of the more recent project sites contained a higher amount of graphics and multimedia than the older ones, due to the increasing technology of the Internet.** About half of the functional projects surveyed (25) contained a project .pdf file. These files are simply digitized plans and descriptions of the project. Nearly all of the functional projects surveyed (44) contained a NSSDC (National Space Science Data Center) site, which are descriptions of the project and project missions, including launch dates and mission objectives.

Only a few (13) included the latest update date. Of those that did, not many were updated regularly. In most instances, the sites had only been updated a few times after the initial construction. This is to be expected because the newer sites are related to ongoing missions and therefore are still being updated, where the other, older sites, are related to missions that have closed and have no more current data to share.

Only 8 of the surveyed functional project sites implemented metatags in their source coding. Of these 8, all of the metatags consisted of, "keyword", "description", or "content" metatags. The keyword or descriptions were normally the name of the project, associated topics, or something dealing with NASA.

Spidering sample project and science sites

Following a review of spidering software, HTTrack Version 3.x, a large-scale spidering and Web site copying program, was downloaded and installed. The spider was run on the Library's Web site, because permission was not needed to perform this spidering. It also helped to familiarize the team with the spidering software and to determine how to customize the spidering software to our needs. The Library's Web site is described below:

Library site: http://library.gsfc.nasa.gov

For the final library spidering, I chose to run the spider to 10 levels down from the Library's homepage, capturing all links and including all domains. The content of the library's site includes: The Goddard Projects Directory, colloquia, books in the library, standards and technical reports, virtual reference shelf, etc. The library's Website is always being updated and added to. This site is very important because it contains the project directory, in which all of the projects are listed, links to project sites and other project information is also given.

The spider was run on the Library Web site three times, each to different levels. The first was to three levels, then to five levels and finally to ten levels. We ran these tests for the purpose of familiarizing ourselves with the HTTrack program and finding out where we need to "tweak" our spider settings so that when we spidered the test sites, we would have the optimal settings for doing so. There are many settings that can be changed before actually running the spider, and each setting has it's own affect on the process. Some of these settings were as follows: Does the spider accept cookies? (set to No), Does the spider follow the robot.txt rules (set to Yes), how many connections to the server do we use, what is the max transfer rate in Bits per second, what is the max number of links we can scan at any given time, what is the total time we want to limit the spidering session to, etc.

Three sites were then selected at random from 2001 in the Project Directory. A letter was sent to the e-mail contacts for those sites in order to tell them that the spider would be run and to inform them about the purpose of the pilot project. The three sites are described below.

MAP Website: http://map.gsfc.nasa.gov/

"MAP is a NASA Explorer Mission that will measure the temperature of the cosmic background radiation over the full sky with unprecedented accuracy. This map of the remnant heat of the Big Bang will provide answers to fundamental questions about the origin and fate of our universe." The content is of course, about a space probe/satellite that measures and records temperature data. Charles Bennett is the NASA official in charge, and is the only email contact provided. The site appears to be updated on a somewhat regular basis, having been updated 07-16-2002".

RHESSI Web site: http://hesperia.gsfc.nasa.gov/hessi/

The HESSI Website is maintained by the Laboratory for Astronomy and Solar Physics, Solar Physics branch code 682. "RHESSI's primary mission is to explore the basic physics of particle acceleration and explosive energy release in solar flares." Contents include: Facts, News, a search function, presentations, related sites, etc. The responsible NASA Official is Gordon D. Holman. The site was last updated on June 13, 2002.

TDRS Web site: http://nmsp.gsfc.nasa.gov/tdrss/tdrshij.html/

The site is about the TDRS program, which consists of a number of satellites launched over the years to perform tests in space. "The Tracking and Data Relay Satellites comprise the space segment of NASA's communications relay system, providing telecommunication services to low earth orbiting spacecraft." There is not a regular schedule for updating the Web site. It is just an archive type Web site that records information with no updates. The responsible NASA official is Jon Walker.

The TDRS site was ultimately excluded from the spidering when it was determined that this site had links to a contractor's server, as well as some external non-science commercial sites. The team replaced this site with the GSFC Technology Page described below.

Technology Page: http://gsfctechnology.gsfc.nasa.gov/

This site is not a project site. The contents include: Current NASA activities, Technology investment areas, Distributed Space Systems, Flight & Science Information Systems, etc. Example data for Distributed Space Systems: "Distributed Space Systems Technology allows NASA to exploit new vantage points, developing new sensing strategies and implementing system-wide techniques which promote agility, adaptability, evolvability, scalability, and affordability through exploitation of multiple space platforms." The site is geared toward explaining some of the more gritty technology used at NASA to the public so they may understand what is being done and how. The Website appears to be updated on an almost daily basis. The responsible NASA official is Lisa Callahan.

The spidering tests ran into a number of different problems. Early on, the problems were merely settings that needed to be tweaked in order to perform at an optimal level. Examples are as follows:

1) **The spider ran out of total time in which to perform the session.** This was fixed by leaving the selection for total time blank, meaning an infinite amount of time, or running until it's finished.
2) **Not going down the correct number of levels.** Fixed also by simply changing the selection for layers scanned.
3) **Spidering down too many levels**. Similar to #2, changed selection for layers to be scanned.
4) **Spidering sites that were not our own was not advisable until the Webmasters/curators of the sites were contacted.**
5) **When the actual project sites were spidered, the spider found so much information that it would not stop spidering.** For the three project sites that were scanned, we took down over 1.5 Gigabytes of information (probably much more). The scan took more than a day and a half, and was nowhere near finished when it was canceled it. For one of the sites, after a day and a half, only 5,333 links of a total of over 40,000 had been scanned. Each time it scanned a new link, it found more, adding to the total number of links each time. One of the movie files that the spider tried to grab was more than 174MB, and the technician finally cancelled it through an override feature included in HTTrack.
6) **Slow equipment used during the pilot was incapable of running multiple spiders at once because it would have been using too much of the systems resources.** This slowed us down a great deal, because the spider had to be scheduled to run at night, each at different times so as not to kill the computer. This meant that the technician was not available to monitor the spidering and correct problems. This spider should be run from a server or a computer dedicated to the task. The computer should have a fast processor and a great deal of storage. A gig and a half was only the first 10% or so, and that has the potential to be only a few percent of a total site spidering. Running the spider on one computer and saving it on another was acceptable for the pilot, but for a production system a dedicated machine(s) will be needed to be most efficient. Toward the end of the pilot period, the computer was replaced with one having a higher speed processor and more storage. This improved the spidering process significantly.

While the spidering software did not provide good statistics with regard to the time required, some statistics were captured for the test sites.

Hessi – Started 15:00:00 Tue. 20 Aug. 2002 ended 11:28:26:19 Thurs. Aug 2002, and it was still not close to completion. User cancelled the process.
MAP - Started 19:00:00, Tue 20 Aug. 2002 ended 09:23:35 Thurs Aug 2002, also not close to complete. User cancelled the process.
Tech Page – Started 20:00:00, Thur. 12 Sep 2002 ended 21:21:33 same day. Site spidering was complete. This site crawling was shorter because it was within the Goddard domain only.

7) **Dealing with non-GSFC sites was identified as a problem during the spidering**. Originally, the spider was set to accept all non-protected sites. This resulted in access to a contractor's computer which is located at GSFC, but is no in the GSFC domain. However, as the problem was investigated, it appeared that someone might have spoofed the IP address in an attempt to hack the contractor's computer. There was more than one instance in which the IP address showed up on the contractor's computer logs when the technician had only run

the spider once. Also, the logs showed that at a different time, another IP address located in the Library tried to log into the contractor's computer as an administrator.

8) **A virus was found on the computer running the spider.** While it isn't clear if the viruses were on the machine before or resulted from the spidering, it raised the possibility of infecting the GSFC system during spidering. This emphasizes the need for an isolated system to handle the spidering. Such a system should be loaded loaded with anti-virus software that is comprehensive and well maintained. Once the content has been checked, it can then be moved over to the archiving system.

Indexing the spidered content

The results of all the spidering, including the Library, project and science Web sites, were provided to the Autonomy Server for full text indexing. This search software provides word and phrase searching. The default Autonomy search interface was modified for this purpose. It is available at http://library01.gsfc.nasa.gov/archive/ (Link to Autonomy Index page). This test included only the HTML and pdf files but did not include any text from metadata that could be associated with pictures or video files.

Developing a pilot semi-automated production environment for Web capture

Toward the end of the pilot period, the team developed a conceptual design for a production system for capturing Web sites. The goal was to build on the findings of the testing done during the pilot period and to automate as much of the process as possible. It also involved not only the spidering and copying of the sites, but the creation of metadata. This work is detailed in Part 2 of this report.

Review of the Spidering Software

While initial analysis of the spidering software available, determine that HTTrack was sufficient for the needs of the project, the actual use of the software provided otherwise. While the general functionality is fairly adequate, there were special needs for statistics that could not be met. The generated log files are nothing more than error log entries, and it is impossible without time consuming manual review to determine the domains of linked sites that are being crawled at various levels. It is impossible to tell at which level a certain entry occured. Since there were eight open connections during spidering, one of those connections might have been faster, and jumped down more levels than another connection, making the log files nothing more than a jumble of error messages without specific regards as to which level they came from. While it may be possible to develop scripts to process the log files, this has not yet been determined. It is anticipated that not only would more detailed statistics be of interest during this research phase, but as an ongoing metric.

HTTrack is a "dumb" program. You can not tell it to copy a website, and expect it to copy only that site. When it jumps a level down, it does not discriminate and will copy all links that are within the parameters, even if they are not part of the main path site. It is especially hard if valid information is linked to the site, but isn't actually part of the project site itself. To make this work, each site would have to have it's own set of specifically tailored parameters, which would not completely eliminate problems.

Therefore, follow-on to this project should include a closer analysis of available spidering/crawling tools and an investigation of the systems that are in use elsewhere by national libraries and other institutions. The team has begun this already by investigating the harvesting system that the National Technical Information Service (NTIS) uses to grab technical reports from the DOE Information Bridge database. While in the past they were using a commercial program, they are now in the process of developing their own.

APPENDIX F
ANALYSIS OF GSFC PROJECT AND NON-PROJECT WEB SITES

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
2001	http://map.gsfc.nasa.gov/	Map	Public, Research	front page, image links, media, search	<meta name="generator" content="GoLive CyberStudio 3">	none		NSSDC site, Additional Link	
2001	http://hesperia.gsfc.nasa.gov/hessi/	RHESSI	Public, Research	Front page, links for related sites, news, software and search functions.	None	none	Jun-02	NSSDC site, Project site, picutre link inoperative	
2000	http://stp.gsfc.nasa.gov/missions/timed/timed.htm%20and%20http://www.timed.jhuapl.edu/	TIMED				403		NSSDC site, Project site (forbidden)	
2000	http://toms.gsfc.nasa.gov/index.html	TOMS	Public, Research	Front page, Multimedia links, links for teachers	None	none	Apr-02	Project site only	
2000	http://image.gsfc.nasa.gov/	Image	Researcher	front page, image links, animation links	<meta name="keywords" content="NASA, GSFC, data, space physics, plasma, magnetosphere, magnetospheric imaging, Imager for Magnetopause-to-Aurora Global Exploration, Explorer program, Goddard Space Flight Center">	none	Jun-02	NSSDC site, Project site	
1999	http://sunland.gsfc.nasa.gov/smex/wire/	Wire	Researcher	images, charts, graphs, page links, data links	<meta name="GENERATOR" content="Microsoft FrontPage 3.0">	none		NSSDC site, Project site	
1999	http://terra.nasa.gov/	TERRA	Public, Research	Front page, many detailed picture links, weekly mission status update,	<META NAME="keywords" CONTENT="Terra,earth science,satellite data images,environment,global change,earth observing system,ASTER,CERES,MISR,MODIS,MOPITT,EOS">	None	Jun-02	NSSDC site, Project site, Project pdf	

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
				images and data links					
1998	http://cfa-www.harvard.edu/swas/	SWAS	Public, Research	Front page, links to image gallery, current status, publications, about	<META NAME="description" CONTENT="The Submillimeter Wave Astronomy Satellite,	None		NSSDC site, Project site, Project pdf, Additional link	
1998	http://lunar.arc.nasa.gov/	Lunar	Public, Research	Front page, data links, archive links, resource links	<meta http-equiv="keywords" content="Lunar Prospector Space Exploration Moon Mission NASA archives photos photographs images altlas scientists project education history">	none		NSSDC site, Project site	
1997	http://goes1.gsfc.nasa.gov/	Goes10				404		NSSDC site, Project site**	
1997	http://seawifs.gsfc.nasa.gov/SEAWIFS.html	SeaWIFS	Public, Research	Front page, data links, image links, resource links, related links	None	none	Jul-02	NSSDC site, Project site, Additional Link	
1997	http://trmm.gsfc.nasa.gov/	TRMM	Public, Research	Front page, good graphics, links to images, movies, publications, related links	<meta name="TRMM Tropical Rainfall Measuring Mission, TRMM" content="TRMM">	none		NSSDC site, Project site, Project pdf	
1996	http://sunland.gsfc.nasa.gov/smex/fast/index.html	FAST	Public, Research	Not many grapshics, lots of technical information, links to latest	None	none	Nov-97	NSSDC site, Project site, Project pdf, Additional link (x2)	

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
				mission info					
1996	http://near.jhuapl.edu/	NEAR	Public, Research	Front page, site map, FAQ, pdf links, movies, search	<meta name="keywords" content="NASA, discovery, asteroid, space, space science, spacecraft, near-earth, NEAR, solar system, planet, planetesimal, multispectral imager, near-infrared spectrograph, flight control, orbit, Eros, 433 Eros, Mathilde">	none	Feb-01	NSSDC site, Project site	
1995	http://goes1.gsfc.nasa.gov/	Goes9				404		NSSDC site, Project site	
1995	http://sohowww.nascom.nasa.gov/	SOHO	Public, Research	Front page, links, search, gallery, about.	<meta http-equiv="Keywords" name="Keywords" content="sun, solar, solar images, helioseismology, solar cor	none		NSSDC site, Project site	
1995	http://rxte.gsfc.nasa.gov/docs/xte/xte_1st.html	RXTE	Public, Research	Front page, links to data archive, analysis, education and outreach	<META Name="keywords" Content="RXTE, XTE GOF, RXTE GOF, XTE, science,	none	Jul-02	NSSDC site, Project site	
1994	http://nssdc.gsfc.nasa.gov/space/istp/wind.html	ISTP	Research	Front page, links to data, space craft info, and instrument info	None	none	Jan-02	NSSDC site, Project site, Project pdf	632
1994	http://nssdc.gsfc.nasa.gov/nmc/tmp/1993-023B.html	Spartan	Public, Research	Info page, little other content	None	none		NSSDC site, Additional Link	
1993	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/tdrs%206.pdf	TDRS F				404		NSSDC site, Project site**, no other projects this year had project sites.	
1993	http://nssdc.gsfc.nasa.gov/nmc/tmp/1993-058C.html	ORFEUS-SPAS 1	Public, Research	Info page, little other content	None	none		NSSDC site, Additional link	
1992	http://nssdc.gsfc.nasa.gov/nmc/tmp/1992-070B.html	LAGEOS II	Public, Research	Info page, little other content	None	none		NSSDC site, Additional link	

APPENDIX F: ANALYSIS OF GSFC PROJECT AND NON-PROJECT WEB SITES

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
1992	http://www-istp.gsfc.nasa.gov/istp/geotail/	Geotail	Public, Research	Front page, no graphics, links to project overview, spacecraft diagram, key parameters	None	none	Sep-01	NSSDC site, Project site	
1992	http://surya.umd.edu/www/sampex.html	SAMPEX	Public, Research	Front page, one graphic, data links, Intro link	None	none		NSSDC site, Project site, Project pdf.	
1991	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/gro.pdf	GRO	Public, Research	None	None	none		Project pdf.	
1991	http://nssdc.gsfc.nasa.gov/nmc/tmp/1991-063B.html	UARS	Public, Research	Info page, little other content	None	none		NSSDC site, Project pdf	
1990	http://heasarc.gsfc.nasa.gov/docs/journal/bbxrt2.html	BBXRT	Research	Front page, Software links, data archive, education and outreach links.	None	none		Project site only	
1990	http://fpd.gsfc.nasa.gov/440/	Hubble	Public, Research	Front page, Hubble links, project links, mission stmt	None	none		NSSDC site, Project site, Project pdf, Additional Link (x2)	440
1989	http://fpd.gsfc.nasa.gov/454/	TDRS D				404		NSSDC site, Project site**, Project pdf, Additional Link	
1989	http://space.gsfc.nasa.gov/astro/cobe/	COBE	Researcher	Front page, Information links, data links	None	none	Jun-01	NSSDC site, Project site, Project pdf.	
1989								***Both projects for this year already recorded	
1988	http://fpd.gsfc.nasa.gov/454/	TDRS C				404		NSSDC site, Project site**, Additional link	

APPENDIX F: ANALYSIS OF GSFC PROJECT AND NON-PROJECT WEB SITES

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
1988	http://crpsm.psm.uniroma1.it/	San Marco	Public, Research	Front page, Information links, data links	None	none		NSSDC site, Project site.	
1987	http://www.hughespace.com/factsheets/376/goes/goes.html	Goes-H	Public, Research	Re-direct to Boeing site map.	None	none		Project site, Additional Link, After re-direct, Goes-H info can be found on Boeing site map.	
1987								All projects sfor this year already recorded	
1986	http://nssdc.gsfc.nasa.gov/nmc/tmp/1986-073A.html	NOAA 10	Public, Research	Info page, little other content	None	none		NSSDC site only	
1986	http://nssdc.gsfc.nasa.gov/nmc/tmp/SPATN-H.html	Spartan-H	Public, Research	Info page, little other content	None	none		NSSDC site, no project this year had a project site.	
1985	http://nssdc.gsfc.nasa.gov/nmc/tmp/1985-048E.html	Spartan-A	Public, Research	Info page, little other content	None	none		NSSDC site, Additional Link	
1984	http://geo.arc.nasa.gov/sge/landsat/landsat.html	Landsat	Public, Research	Front page, image gallery, news, project summary	None	none	Oct-99	NSSDC site, Project site.	
1984	http://sd-www.jhuapl.edu/AMPTE/	AMPTE	Research	Front page, links to images, publications and presentations, data archives	None	none	Mar-99	NSSDC site, Project site, Additional Link	
1983	http://nssdc.gsfc.nasa.gov/nmc/tmp/1983-026B.html	TDRS A	Public, Research	Info page, little other content	None	none		NSSDC site, Project pdf	
1983	http://goes1.gsfc.nasa.gov/	Goes-F				404		NSSDC site, Project site**	
1983	http://nssdc.gsfc.nasa.gov/nmc/tmp/1983-022A.html	NOAA 8	Public, Research	Info page, little other content	None	none		NSSDC site, no other project this year had project site.	

APPENDIX F: ANALYSIS OF GSFC PROJECT AND NON-PROJECT WEB SITES

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
1982	http://nssdc.gsfc.nasa.gov/nmc/tmp/1982-022A.html	STS3/ OSS1	Public, Research	Info page, little other content	None	none		NSSDC site, Additional Link, no project this year had a project site.	
1982	http://nssdc.gsfc.nasa.gov/nmc/tmp/1982-072A.html	Landsat D	Public, Research	Info page, little other content	None	none		NSSDC site, Project pdf, Additional link	
1981	http://nssdc.gsfc.nasa.gov/nmc/tmp/1981-100A.html	SME	Public, Research	Info page, little other content	None	none		NSSDC site only	
1981	http://nssdc.gsfc.nasa.gov/nmc/tmp/1981-049A.html	Goes 5	Public, Research	Info page, little other content	None	none		NSSDC site, no other project this year had project site.	
1980	http://nssdc.gsfc.nasa.gov/nmc/tmp/1980-014A.html	SMM	Public, Research	Info page, little other content	None	none		NSSDC site, Project pdf, Additional Link, no other project this year had a project site.	
1980	http://nssdc.gsfc.nasa.gov/nmc/tmp/1980-074A.html	Goes D	Public, Research	Info page, little other content	None	none		NSSDC site, Project pdf	
1979	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/magsat.pdf	MAGSAT	Public, Research	Info page, little other content	None	none		Project pdf only	
1979	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/sage.pdf	SAGE	Public, Research	Info page, little other content	None	none		Project pdf, no other project this year had a project site.	
1978	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/iue.pdf	IUE	Public, Research	Info page, little other content	None	none		Project pdf, no other project this year had a project site.	
1978	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/tiros%20n.pdf	Tiros N				404		Project pdf only	
1978	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/heao2.pdf	HEAO 2	Public, Research	Info page, little other content	None	none		Project pdf only	
1977	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/isee2.pdf	ISEE 2	Public, Research	Info page, little other content	None	none		Project pdf only	
1977	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/goes%202.pdf	Goes/NOAA				404		Project pdf	
1977	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/heao1.pdf	HEAO 1	Public, Research	Info page, little other content	None	none		Project pdf, no other project this year had a project site.	
1976								There were no projects this year that had project sites or NSSDC sites,	

Year	URL	Descr.	Audience	Link Types	Use of meta tags/which	Probs.	Up-date	Notes	Code
								only single pictures.	
1976								There were no projects this year that had project sites or NSSDC sites.	
1975	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/goes%201.pdf	SMS-C				404		Project pdf only**	
1975	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/sms%202.pdf	SMS-B				404		Project pdf only ** There were no more projects this year with links or sites.	
1975	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/landsat2.pdf	Landsat2	Public, Research	Info page, little other content	None	none		No sites this year had project sites, only NSSDC sites.	
1974	http://library.gsfc.nasa.gov/GdrdProjs/ProjInfo/sms%201.pdf	SMS A				404		Project pdf, no other project this year had a project site.	
1959 – 1974								No other sites this year had project site, and no NSSDC sites.	

NON PROJECT SITES

URL	Descr.	Audience	Link Types	Use of meta tags/which	Problems	Update	Notes	Code	
http://gsfctechnology.gsfc.nasa.gov/	Tech. @ Goddard	Public	Links to different systems, current activities, mission infusion, technology investments	none	none	Aug-02	Link from main GSFC page		a single opening graphic, links lead to many more
http://newsmedia.gsfc.nasa.gov/	News/media	Public	Links to different stories in the news about what is going on at gsfc or what gsfc is doing	none	none	Aug-02	Link from main GSFC page		Many thumbnail graphics linked to larger and more graphics
http://www.gsfc.nasa.gov/earth.html	Earth	Public	Links to earth science missions and other earth science related links	none	none		Link from main GSFC page	295/293	Background pic only
http://www.gsfc.nasa.gov/space.html	Space	Public	Link to site dealing with space science missions	none	none		Link from main GSFC page	295/293	Background pic only
http://www.gsfc.nasa.gov/mission.html	Missions	Public	Links to the different missions, missions to launch and launched missions.	none	none		Link from main GSFC page	295/293	Background pic only
http://www.gsfc.nasa.gov/photos.html	Photos	Public	Links to many different pictures galleries, sattelite images etc.	none	none		Link from main GSFC page	295/293	Background pic only
http://www.gsfc.nasa.gov/public.html	Public info	Public	Links to different colloquias and seminars, goddard news and news archives	none	none		Link from main GSFC page	295/294	Background pic only
http://www.gsfc.nasa.gov/topstory/20020722landsat30.html	LandSat art	Public	Links to many different and interesting animations as seen from LandSat	none	none	Jul-02	Link from main GSFC page		Background pic, thumbnail pics of the animations

NON PROJECT SITES									
URL	**Descr.**	**Audience**	**Link Types**	**Use of meta tags/which**	**Problems**	**Update**	**Notes**	**Code**	
http://www.nasa.gov/today/index.html	Today	Public	Links to recent news releases and other current information		none	Aug-02	Link from main GSFC page		thumbnail pictures of selectable topics
http://earthobservatory.nasa.gov/NaturalHazards/natural_hazards_v2.php3?img_id=4654	Fire	Public	Links to images dealing with the gallery, features, data, reference	none	none		Link from main GSFC page		Large single pic, links to many others.

APPENDIX G
Conversion of HTML To Dublin Core Metatags and XML

Samples of metadata automatically generated from three GSFC project homepages. First example is the Dublin Core metadata, which can be submitted back as HTML metatags. The second example in each set is the equivalent content encoded in XML, which would be appropriate for submission to a database as part of the preservation/descriptive metadata.

Sample 1: Generated metadata HTML code for MAP Website(homepage)

<link rel="schema.DC" href="http://purl.org/dc">
<meta name="DC.Title" content="Microwave Anisotropy Probe - Cosmology">
<meta name="DC.Subject" content="It appears that you do not have Javascript enabled on your browser, or you have a browser version older than 4.0. The site can be navigated without these, but the experience is better with a 4.0 or greater browser with Javascript turned on; Charles L. Bennett / Charles.L.Bennett.1@gsfc.nasa.gov; If you need to upgrade your browser follow these links; CONTINUE">
<meta name="DC.Type" scheme="DCMIType" content="Text">
<meta name="DC.Format" content="text/html">
<meta name="DC.Format" content="25057 bytes">
<meta name="DC.Identifier" content="http://map.gsfc.nasa.gov/">

Sample 1: Generated XML metadata for MAP Website(homepage)

<?xml version="1.0" ?>

- <metadata xmlns:dc="http://purl.org/dc/elements/1.1/">

<dc:title>Microwave Anisotropy Probe - Cosmology</dc:title>

<dc:subject>It appears that you do not have Javascript enabled on your browser, or you have a browser version older than 4.0. The site can be navigated without these, but the experience is better with a 4.0 or greater browser with Javascript turned on; Charles L. Bennett / Charles.L.Bennett.1@gsfc.nasa.gov; If you need to upgrade your browser follow these links; CONTINUE</dc:subject>

<dc:type>Text</dc:type>

<dc:format>text/html || 25057 bytes</dc:format>

<dc:identifier>http://map.gsfc.nasa.gov/</dc:identifier>

</metadata>

Sample 2: Generated HTML metadata for HESSI Website(homepage)

<link rel="schema.DC" href="http://purl.org/dc">
<meta name="DC.Title" content="RHESSI Home Page">
<meta name="DC.Subject" content="Responsible NASA Official; ; This site last updated June 13, 2002; Web Sites; holman@stars.gsfc.nasa.gov; Laboratory for Astronomy and Solar Physics; Small Explorer; Merrick Berg, Brian Dennis, Gordon Holman; RHESSI was launched on February 5, 2002; ; Major Events; RHESSI is a NASA; Other RHESSI; Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI; Web Design; Welcome to NASA Goddard's home page for the; RHESSI's primary mission is to explore the basic physics of particle acceleration and explosive energy release in solar flares; Gordon D. Holman; First-Light Press Release; Gilbert Prevost; June 5 Press Release">
<meta name="DC.Date" scheme="W3CDTF" content="2002-07-10">
<meta name="DC.Type" scheme="DCMIType" content="Text">
<meta name="DC.Format" content="text/html">
<meta name="DC.Format" content="27233 bytes">
<meta name="DC.Identifier" content="http://hesperia.gsfc.nasa.gov/hessi/">

Sample 2: Generated XML metadata for HESSI Website(homepage)

<?xml version="1.0" ?>

- <metadata xmlns:dc="http://purl.org/dc/elements/1.1/">

<dc:title>RHESSI Home Page</dc:title>

<dc:subject>Responsible NASA Official; ; This site last updated June 13, 2002; Web Sites; holman@stars.gsfc.nasa.gov; Laboratory for Astronomy and Solar Physics; Small Explorer; Merrick Berg, Brian Dennis, Gordon Holman; RHESSI was launched on February 5, 2002; ; Major Events; RHESSI is a NASA; Other RHESSI; Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI; Web Design; Welcome to NASA Goddard's home page for the; RHESSI's primary mission is to explore the basic physics of particle acceleration and explosive energy release in solar flares; Gordon D. Holman; First-Light Press Release; Gilbert Prevost; June 5 Press Release</dc:subject>

<dc:date>2002-07-10</dc:date>

<dc:type>Text</dc:type>

<dc:format>text/html || 27233 bytes</dc:format>

<dc:identifier>http://hesperia.gsfc.nasa.gov/hessi/</dc:identifier>

</metadata>

Sample 3: Generated HTML metadata for TDRSS Website

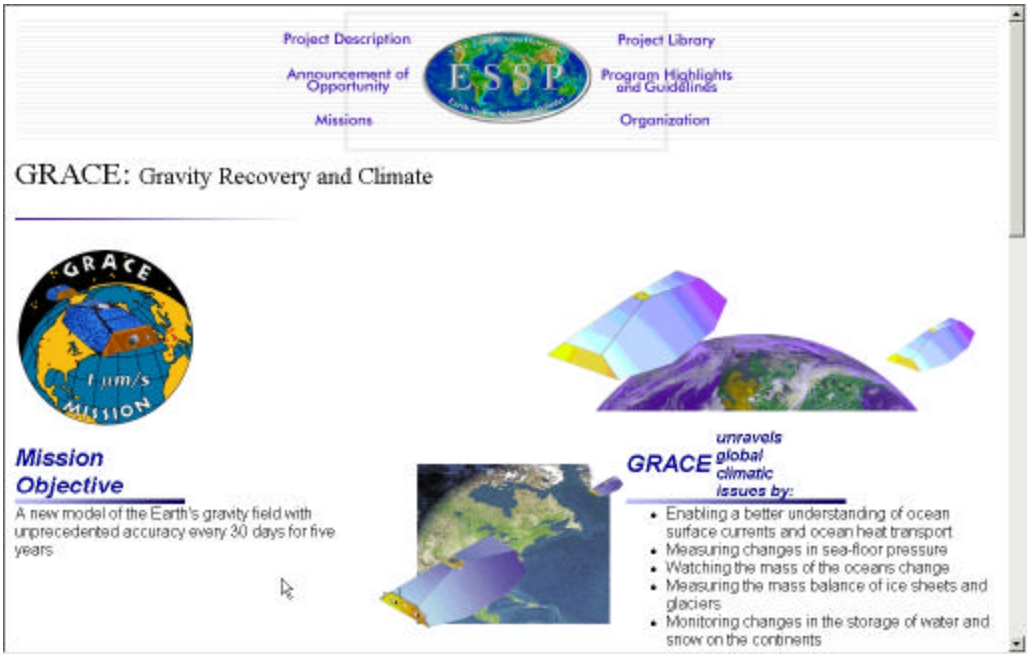
<link rel="schema.DC" href="http://purl.org/dc">
<meta name="DC.Title" content="TDRS H, I, J The Next Generation">
<meta name="DC.Subject" content="Spacecraft Characteristics; Service Comparison between TDRS H, I, J and F-1 through F-7 (original design; TDRS H, I, J Program Requirements; Return to Spacecraft Homepage; Ka-Band Features">
<meta name="DC.Date" scheme="W3CDTF" content="2000-05-25">
<meta name="DC.Type" scheme="DCMIType" content="Text">

```
<meta name="DC.Format" content="text/html">
<meta name="DC.Format" content="879 bytes">
<meta name="DC.Identifier" content="http://nmsp.gsfc.nasa.gov/tdrss/tdrshij.html">
```

Sample 3: Generated XML metadata for TDRSS Website

```
<?xml version="1.0"?>
<metadata
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>
    TDRS H, I, J The Next Generation
  </dc:title>
  <dc:subject>
    Spacecraft Characteristics; Service Comparison between TDRS
    H, I, J and F-1 through F-7 (original design; TDRS H, I, J
    Program Requirements; Return to Spacecraft Homepage; Ka-
    Band Features
  </dc:subject>
  <dc:date>
    2000-05-25
  </dc:date>
  <dc:type>
    Text
  </dc:type>
  <dc:format>
    text/html || 879 bytes
  </dc:format>
  <dc:identifier>
    http://nmsp.gsfc.nasa.gov/tdrss/tdrshij.html
  </dc:identifier>
</metadata>
```

APPENDIX H
Examples of GSFC Web Pages

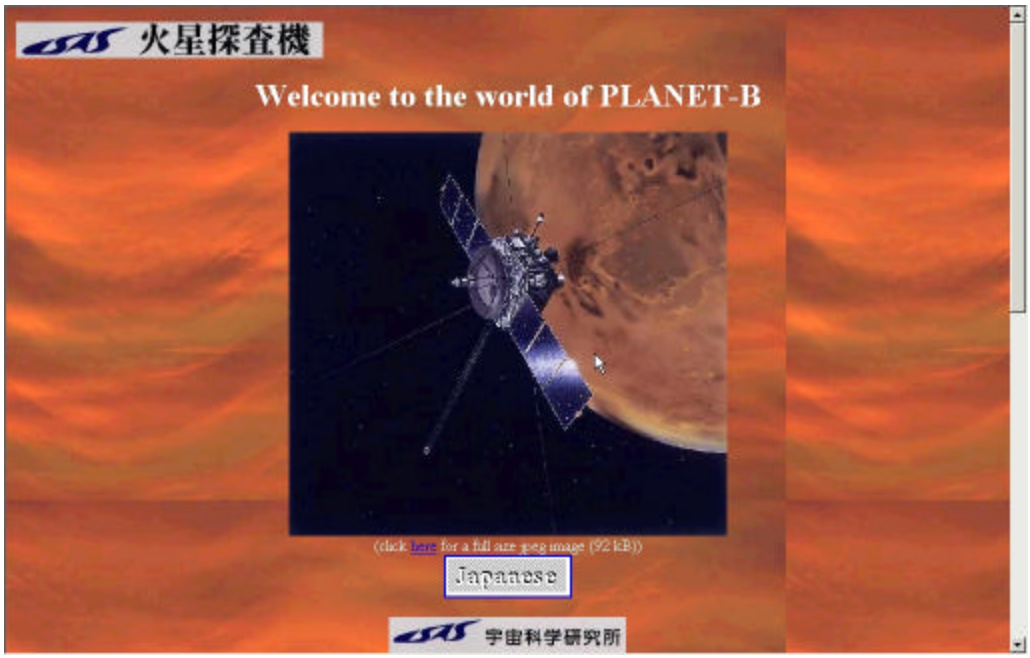


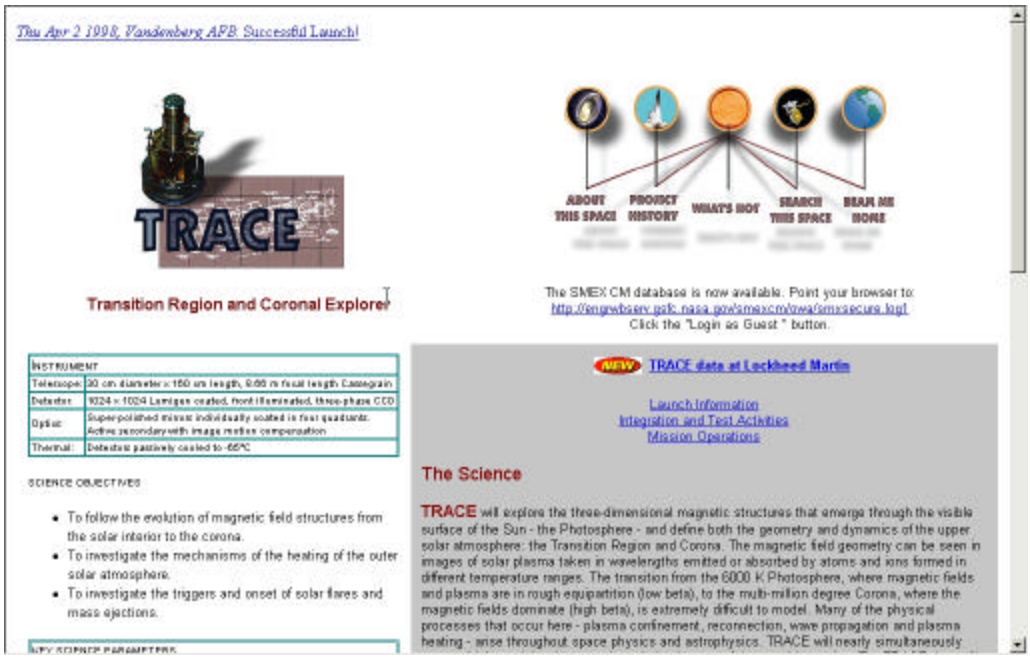

Tracking and Data Relay Satellite H, I, J
The Next Generation
TDRS H, I, J Program Requirements
Spacecraft Characteristics
Ka-Band Features
Service Comparison between TDRS H, I, J and F-1 through F-7 (original design)
Return to Spacecraft Homepage

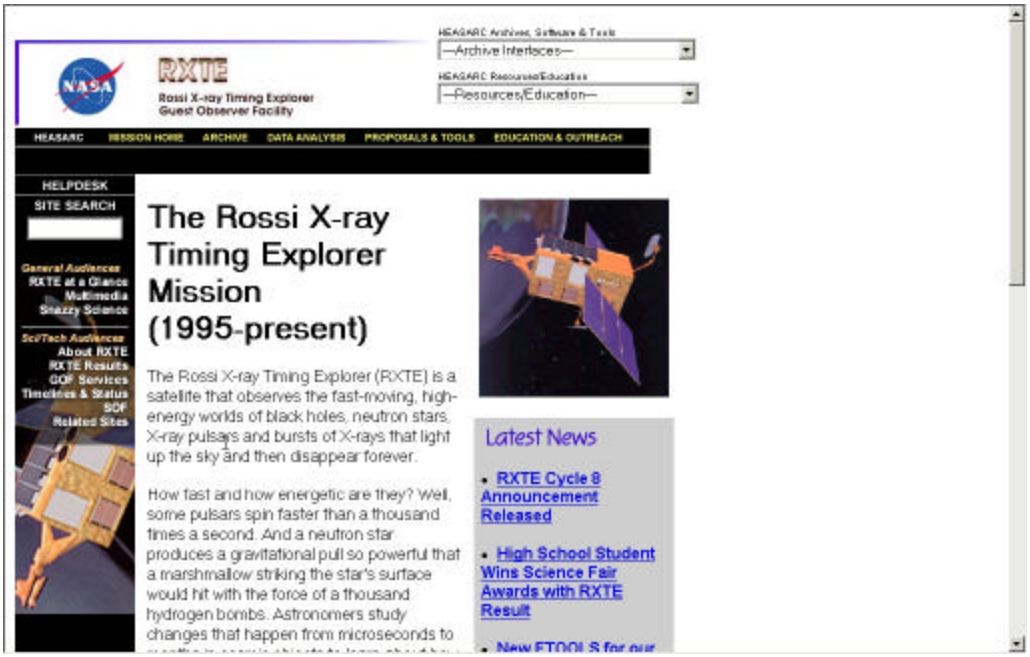

MAP is a NASA Explorer mission that will measure the temperature of the cosmic background radiation over the full sky with unprecedented accuracy. This map of the remnant heat from the Big Bang will provide answers to fundamental questions about the origin and fate of our universe.
Microwave Anisotropy Probe
MAP Mission
Help/Search
Outreach/Media
Image Gallery
NASA / GSFC / Princeton / Privacy & Security / Mission Status


The Submillimeter Wave Astronomy Satellite
(SWAS)
Menu
SWAS Home
The Mission
The Instrument
The Spacecraft
Current Status
SWAS Science
Publications
Gallery
People
Links
Welcome to the SWAS Home Page
SWAS is a NASA Small Explorer Project (SMEX) designed to study the chemical composition of interstellar gas clouds. SWAS was launched into low Earth orbit on December 05, 1998. The primary objective of SWAS is to survey water, molecular oxygen, carbon, and isotopic carbon monoxide emission in a variety of galactic star forming regions. The SWAS Science Operations Center is located at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.


Welcome to the Moon!
The Lunar Prospector Mission A Success
The NASA/Ames Research Center in Mountain View, CA invites you and the world to look at the events surrounding the first NASA Moon mission in 25 years!
"Lunar Prospector has given us new eyes to look at the solar system"
lunar prospector
latest news
credit and contacts
project results datavis history science education archives resources
To request information on this web site in a Section 508 accessible format, please contact access@mail.arc.nasa.gov


火星探査機
Welcome to the world of PLANET-B
(click here for a full size jpeg image (92 kB))
Japanese
宇宙科学研究所

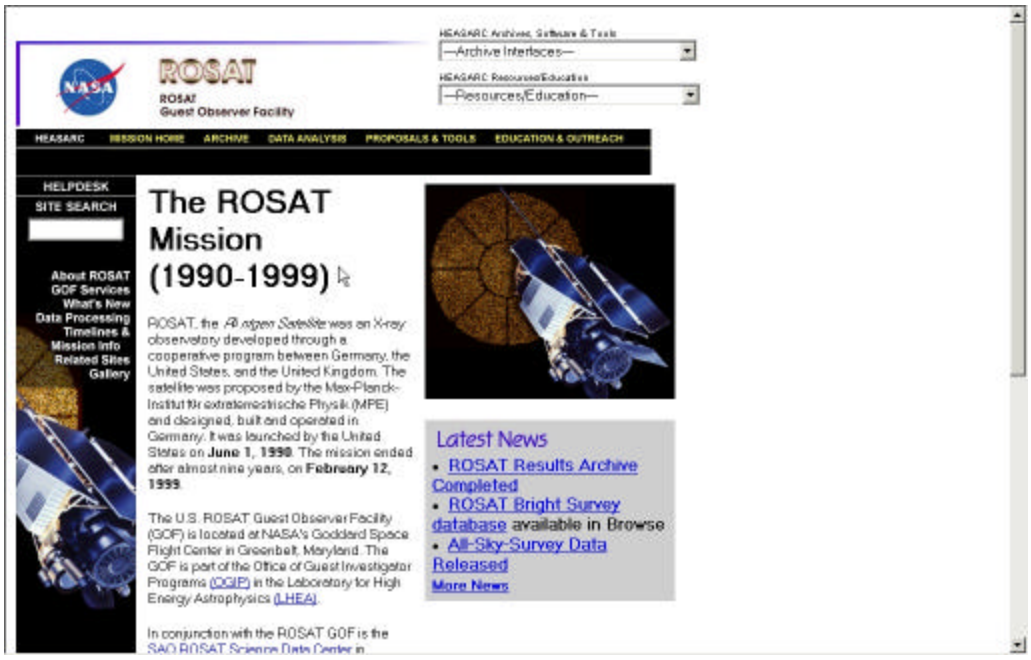

Thu Apr 2 1998, Vandenberg AFB: Successful Launch!
TRACE
Transition Region and Coronal Explorer
ABOUT THIS SPACE
PROJECT HISTORY
WHAT'S HOT
SEARCH THIS SPACE
BEAM ME HOME
The SMEX CM database is now available. Point your browser to:
http://engwebserv.gsfc.nasa.gov/smexcm/owa/smx.secure_login
Click the "Login as Guest " button.
NEW TRACE data at Lockheed Martin
Launch Information
Integration and Test Activities
Mission Operations
INSTRUMENT
Telescope: 30 cm diameter x 150 cm length, 8.66 m focal length Cassegrain
Detector: 1024 x 1024 Lumigen coated, front illuminated, three-phase CCD
Optics: Super-polished mirrors individually coated in four quadrants. Active secondary with image motion compensation
Thermal: Detectors passively cooled to -66°C
SCIENCE OBJECTIVES
To follow the evolution of magnetic field structures from the solar interior to the corona.
To investigate the mechanisms of the heating of the outer solar atmosphere.
To investigate the triggers and onset of solar flares and mass ejections.
The Science
TRACE will explore the three-dimensional magnetic structures that emerge through the visible surface of the Sun - the Photosphere - and define both the geometry and dynamics of the upper solar atmosphere: the Transition Region and Corona. The magnetic field geometry can be seen in images of solar plasma taken in wavelengths emitted or absorbed by atoms and ions formed in different temperature ranges. The transition from the 6000 K Photosphere, where magnetic fields and plasma are in rough equipartition (low beta), to the multi-million degree Corona, where the magnetic fields dominate (high beta), is extremely difficult to model. Many of the physical processes that occur here - plasma confinement, reconnection, wave propagation and plasma heating - arise throughout space physics and astrophysics. TRACE will nearly simultaneously


RXTE
Rossi X-ray Timing Explorer
Guest Observer Facility
HEASARC
MISSION HOME
ARCHIVE
DATA ANALYSIS
PROPOSALS & TOOLS
EDUCATION & OUTREACH
HELPDESK
SITE SEARCH
The Rossi X-ray Timing Explorer Mission (1995-present)
The Rossi X-ray Timing Explorer (RXTE) is a satellite that observes the fast-moving, high-energy worlds of black holes, neutron stars, X-ray pulsars and bursts of X-rays that light up the sky and then disappear forever.
How fast and how energetic are they? Well, some pulsars spin faster than a thousand times a second. And a neutron star produces a gravitational pull so powerful that a marshmallow striking the star's surface would hit with the force of a thousand hydrogen bombs. Astronomers study changes that happen from microseconds to
Latest News
RXTE Cycle 8 Announcement Released
High School Student Wins Science Fair Awards with RXTE Result


September 05, 2002 17:51:16 UT - Mission Day: 2470 - DOY: 248
HOT SHOTS: The Sun as Art
Weekly Pick: Sunspot cycle - on the downward slope (Sept 5, 2002)
SOHO-500 Comet Contest -- We have a Winner!
SOHO
EXPLORING THE SUN
esa
NASA
SOHO: The Solar and Heliospheric Observatory
SOHO is a project of international cooperation between ESA and NASA
Textonly Version - European Site - US Site
SUNSPOTS
SPACE WEATHER
Estimated Kp
SOLAR WIND
At 17:30 UT
Speed:
447 km/s
Density:
5.80 p/cm³


HEASARC
High Energy Astrophysics
Science Archive Research Center
HEASARC Archives, Software & Tools:
—Archive Interfaces—
HEASARC Resources/Education:
—Resources/Education—
HEASARC
Observatories
Archive
Software
Tools
Help/FAQ
Education & Outreach
HEASARC Services
Information
What's New
Links
Calibration
Images
Site Map
Dedicated Support Facilities
ASCA
Astro-E2
BeppoSAX
CGRO
Chandra
EUVE
HETE-2
Integral
ROSAT
RXTE
XMM-Newton
Other NASA Archive Services
ADS
IRSA
MAST
NED
NSSDC
Other Services
CDS
Hipparcos
ISO
Preprints
Search the HEASARC site:
enter text
The BBXRT Archive
Alan P. Smale
LHEA
1 Overview
The Broad Band X-ray Telescope (BBXRT) was flown on the space shuttle Columbia (STS-35) on 1990 December 2-December 11, as part of the ASTRO-1 payload. The flight of BBXRT marked the first opportunity for performing X-ray observations over a broad energy range (0.3-12 keV) with a moderate energy resolution (typically 90 eV and 150 eV at 1 and 6 keV respectively). This energy resolution coupled with an extremely low detector background made BBXRT a very powerful tool for the study of continuum and line emission from cosmic sources, and the scientific results form a tasty appetizer of what can be expected from Astro-D. The observing program was designed to be an even mix of Galactic and extragalactic targets, although the Galactic Center region was not available due to the time of year BBXRT was launched.
BBXRT was designed and built at the Laboratory for High Energy Astrophysics at


Hubble Space Telescope - Code 440
Home
Mission Statement
Projects
HST Observatory
Management
Links


ROSAT
ROSAT
Guest Observer Facility
HEASARC Archives, Software & Tools
—Archive Interfaces—
HEASARC Resources/Education
—Resources/Education—
HEASARC
MISSION HOME
ARCHIVE
DATA ANALYSIS
PROPOSALS & TOOLS
EDUCATION & OUTREACH
HELPDESK
SITE SEARCH
About ROSAT
GOF Services
What's New
Data Processing
Timelines &
Mission Info
Related Sites
Gallery
The ROSAT Mission (1990-1999)
ROSAT, the Röntgen Satellite was an X-ray observatory developed through a cooperative program between Germany, the United States, and the United Kingdom. The satellite was proposed by the Max-Planck-Institut für extraterrestrische Physik (MPE) and designed, built and operated in Germany. It was launched by the United States on June 1, 1990. The mission ended after almost nine years, on February 12, 1999.
The U.S. ROSAT Guest Observer Facility (GOF) is located at NASA's Goddard Space Flight Center in Greenbelt, Maryland. The GOF is part of the Office of Guest Investigator Programs (OGIP) in the Laboratory for High Energy Astrophysics (LHEA).
In conjunction with the ROSAT GOF is the
SAO ROSAT Science Data Center in
Latest News
• ROSAT Results Archive Completed
• ROSAT Bright Survey database available in Browse
• All-Sky-Survey Data Released
More News